THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING

ASPECTS OF ELECTRICAL COMMUNICATION

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM
TECHNICAL JOURNAL

Distribution of the Duration of Fades in Radio Transmission:

Gaussian Noise Model

By S. O. RICE

(Manuscript received November 29, 1957)

The fluctuations of a received radio signal due to fading are assumed to behave like the envelope of narrow-band Gaussian noise. Estimates of the distribution of the fade lengths for various depths of fades are given, and relations which may be useful in analyzing fading data are derived. A similar problem involving the separation of the intercepts of the noise current itself, instead of its envelope, is also discussed.

I. INTRODUCTION

When a radio transmitter sends out a steady sine wave of very high frequency the signal received far beyond the horizon is no longer steady. The envelope of the received signal fluctuates in an irregular manner and may go through several maxima and minima in a second. These fluctuations have been studied by many people.* For observations extending over fifteen minutes or so, the variations in the size of the envelope follow, approximately, a Rayleigh distribution. For observations extending over a month, the distribution of the hourly medians is roughly normal in decibels with a standard deviation of 8 or 10 db. There are also seasonal and diurnal variations.

---

* In this paper we shall make use of the data given by K. Bullington, W. J. Inkster and A. L. Durkee.[1]

581

Here we shall be concerned mostly with the short-term Rayleigh fading. In particular we shall be interested in the distribution of the lengths of fades, i.e., in the lengths of the intervals during which the signal remains below some specified level — the lower the level, the "deeper" the fade.

We shall assume, in our calculations, that the envelope of the received wave is the same as the envelope of a Gaussian random noise current having a narrow, normal-law power spectrum centered on the transmitter frequency.* Under suitable conditions the distribution of the fade lengths approaches a limiting form for deep fades and the exact form of the power spectrum becomes unimportant.

Although the results obtained from the Gaussian noise model do not always agree with the rather meager available experimental data, they do show general trends. Moreover, a process of averaging over the parameters given in Section IV suggests relations between the depths of fades and their durations which may be of use in analyzing fading data.

Our method of studying the duration of fades may be described as follows. We first find an expression which approximates the distribution function in the region of short intervals. This is done by using the method developed in Section 3.4 of Ref. 2 for dealing with the intervals between the zeros of a random noise current. The second step is concerned with the very long intervals. Although it seems to be generally accepted that the probability of a fade exceeding a great length T decreases as $\exp(-\alpha T)$, no convenient method of calculating the exact value of α is available as yet. Here we more or less ignore this difficulty by using a method which amounts to plotting, on suitable coordinates, our approximation for short intervals and then extrapolating by eye so as to obtain a "reasonable" looking curve. This method is described in Section VI. It turns out that the slope of the extrapolated curve at the origin gives the value of α. Needless to say, the reader should regard these portions (the long fade portions) of our distribution curves merely as estimates.

In Section II we state some known results regarding the frequency and average duration of fades based on the Gaussian noise model. Section III lists the main results obtained in the later sections. In Section IV some of the results set forth in Section III are applied to radio fading problems and comparisons with experimental data are given. Section V is concerned with the lengths of the intervals during which a

---

* This power spectrum is obtained when the propagation is assumed to result from randomly moving scatterers having a Maxwellian distribution of velocities. Even if the power spectrum is not normal most of our formulas will still hold, but some of our curves will no longer apply.

Gaussian noise current $I(t)$ exceeds some given value I. Although the material in Section V is not directly related to the fading of the envelope, it is of interest in its own right and serves to set the stage for the work of Section VII, which deals with the intervals during which the envelope $R(t)$ of $I(t)$ lies below some given value R. A discussion of the computations is given in Section VIII. Several known results concerning the separation of the zeros of a simple Markoff process are listed in Section IX and their relation to the present work is pointed out.

II. THE AVERAGE DURATION OF A FADE

Here we shall state several known results regarding the fluctuation of a Gaussian noise current $I(t)$ and its envelope $R(t)$.* Let the mean square value of $I(t)$ be b (b is the same as ψ_0 and b_0 of Refs. 2 and 3). Then

$$\text{probability } I(t) > I = \tfrac{1}{2}[1 - P(Ib^{-1/2})], \tag{1}$$

$$\text{probability } R(t) > R = e^{-R^2/2b}, \tag{2}$$

where $P(x)$ is the error integral

$$P(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt. \tag{3}$$

Let $w(f)$ be the power spectrum of $I(t)$ and β be the $-\psi_0''$ of Ref. 2 so that

$$\beta = \int_0^\infty w(f)(2\pi f)^2\, df. \tag{4}$$

Then, according to a result due to Mark Kac, the expected number of times per second $I(t)$ passes upward across the level $I(t) = I$ is

$$N_I = \frac{1}{2\pi}\,(\beta/b)^{1/2} e^{-I^2/2b}. \tag{5}$$

The average length of the intervals during which $I(t) > I$ is the quotient of (1) and (5):

$$\bar{t} = \pi(b/\beta)^{1/2} e^{I^2/2b}[1 - P(Ib^{-1/2})],$$

$$\bar{t} \sim \frac{b}{I}\,(2\pi/\beta)^{1/2}, \qquad\qquad I \gg b^{1/2} \tag{6}$$

$$\bar{t} \sim 2\pi(b/\beta)^{1/2} e^{I^2/2b}, \qquad\qquad I \ll -b^{1/2}$$

---

* See Refs. 2 and 3, where reference is made to the earlier work of V. D. Landon, K. A. Norton and others.

where the first asymptotic value follows from

$$\tfrac{1}{2}[1 - P(x)] \sim \frac{e^{-x^2/2}}{x\sqrt{2\pi}}, \qquad\qquad x \gg 1. \quad (7)$$

The corresponding results for the envelope run as follows.* Let $w(f)$ be confined to a narrow band and be symmetrical about the midband frequency f_0, and let β_r be the b_2 of Ref. 3 so that

$$\beta_r = 4\pi^2 \int_0^\infty w(f)(f - f_0)^2 \, df. \qquad (8)$$

Then the expected number of times per second $R(t)$ passes downward across the level R is

$$N_R = (\beta_r/2\pi)^{1/2}(R/b)e^{-R^2/2b}. \qquad (9)$$

The average length of the intervals during which $R(t) < R$ is, from (2) and (9),

$$\bar{t} = (b/R)\,(2\pi/\beta_r)^{1/2}(e^{R^2/2b} - 1). \qquad (10)$$

When R becomes small

$$\bar{t} \to (R/2)\,(2\pi/\beta_r)^{1/2}. \qquad (11)$$

In most of the following work we shall set b equal to 1 so that

$$\overline{I^2(t)} = b = \int_0^\infty w(f) \, df = 1. \qquad (12)$$

When this is done we have

$$\begin{aligned}
\bar{t}/\bar{t}_0 &= [1 - P(I)] \exp (I^2/2), \\
N_I/N_0 &= \exp (-I^2/2),
\end{aligned} \qquad (13)$$

where N_0 denotes the average number of upward crossings, per second, across the axis $I(t) = 0$ and $\bar{t}_0$ the average separation of the zeros of $I(t)$. Table I gives values of these ratios. They do not depend upon the power spectrum $w(f)$.

The ratios corresponding to (13) for $R(t)$ are

$$\begin{aligned}
\bar{t}/\bar{t}_1 &= (e^{R^2/2} - 1)/R(e^{1/2} - 1) \approx 1.54\,(e^{R^2/2} - 1)/R, \\
N_R/N_1 &= R \exp [(1 - R^2)/2] \approx 1.649 \, Re^{-R^2/2},
\end{aligned} \qquad (14)$$

---

* K. A. Norton and his colleagues[4] have considered more general fading problems in which the received wave contains both a steady sinusoidal component and a Rayleigh-distributed component.

TABLE I — AVERAGE DURATIONS AND
CROSSING RATES — $I(t)$

$I/\text{rms } I(t)$	$\bar{t}/\bar{t}_0$	N_I/N_0	$N_I\bar{t}$
−5	536,674	3.73×10^{-6}	$1 - 2.87 \times 10^{-7}$
−4	5,962	3.35×10^{-4}	0.999968
−3	180	0.0111	0.9986
−2	14.4	0.135	0.977
−1	2.77	0.607	0.841
0	1.000	1.000	0.500
1	0.523	0.607	0.159
2	0.336	0.135	0.0228
3	0.243	0.0111	0.00135
4	0.189	3.35×10^{-4}	3.17×10^{-5}
5	0.154	3.73×10^{-6}	2.87×10^{-7}

where $\bar{t}_1$ and N_1 refer to the level $R = 1$. Other reference levels sometimes used are

$$\text{median } R = (2 \log 2)^{1/2} = 1.1774 \cdots \approx 1.18,$$

$$\text{average } R = (\pi/2)^{1/2} \quad \approx 1.25, \qquad (15)$$

$$\text{rms } R = 2^{1/2} \quad \approx 1.41.$$

Thus, instead of the second of equations (14), it is sometimes convenient to use

$$\frac{N_R}{N_m} = \frac{2R}{R_m} e^{-R^2/2} = \frac{2R}{R_m} \left(\frac{1}{2}\right)^{R^2/R_m^2},$$

where N_m and R_m refer to the median level. In (14) and (15) the value of b is unity, i.e., rms $I(t) = 1$, $I(t)$ being the noise current whose envelope is $R(t)$. Table II gives values of the ratios (14). Just as in the case of $I(t)$, these ratios do not depend on the power spectrum.

TABLE II — AVERAGE DURATIONS AND
CROSSING RATES — $R(t)$

$R/\text{rms } I(t)$	$\bar{t}/\bar{t}_1$	N_R/N_1	$N_R\bar{t}$
0.125	0.0967	0.2045	0.00778
0.25	0.1957	0.3995	0.0308
0.5	0.4105	0.7275	0.1175
1.0	1.000	1.000	0.3935
1.18 (median)	1.309	0.9731	0.5000
2.0	4.924	0.4462	0.8647
3.0	45.74	0.0550	0.9889
4.0	1148	0.002212	0.999665

III. GAUSSIAN NOISE RESULTS — NORMAL-LAW POWER SPECTRUM

Let $I(t)$ be a Gaussian noise current with the power spectrum

$$w(f) = \frac{2}{\sigma \sqrt{2\pi}} e^{-f^2/2\sigma^2} \tag{16}$$

so that the mean square value of $I(t)$ is $b = 1$. When $I(t)$ crosses upward through the level I the probability that it will remain above I for more than τ seconds may be denoted by $F(u, I)$. Here

$$u = \tau/\bar{t},$$
$$\bar{t} = (2\sigma)^{-1} [1 - P(I)] \exp (I^2/2), \tag{17}$$
$$\beta = (2\pi\sigma)^2,$$

where $\bar{t}$ is the average length obtained from (6).

Fig. 1 shows our estimates of $F(u, I)$ plotted against u for various values of I. These curves were obtained by the calculations described in Section VIII. As mentioned in the Introduction, the values of $F(u, I)$ for the larger values of u must be regarded as conjectures. Two special cases are

$$F(u, I \to -\infty) = e^{-u},$$
$$F(u, I \to \infty) = e^{-u^2 \pi/4}, \tag{18}$$

and these hold not only for (16) but apparently for any power spectrum such that $F(u, I)$ exists. The gradual change of $F(u, I)$ from an exponential form to a Rayleigh form as I runs from $-\infty$ to $+\infty$ [as indicated by (18)] was surmised by D. S. Palmer[5] from an examination of some experimental data.*

A case which has received some attention in the literature occurs when

$$w(f) = 8/[1 + (2\pi f)^2]^2. \tag{19}$$

This power spectrum is unusual because it represents a borderline case between a simple Markoff process, for which $F(u, I)$ does not exist, and the smoother noise processes, one of which is represented by the $w(f)$ of (16). An experimental investigation of the spacings between zeros of $I(t)$ when $w(f)$ has the form (19) [and also when $w(f)$ has several similar forms] has been carried out by R. R. Favreau, H. Low and I. Pfeffer.[6]

---

* Mr. Palmer has informed me that the "Note added in proof" appearing in his paper should be disregarded. He has obtained a proof that the Rayleigh form is indeed approached at $I = \infty$. His proof is given in our Section V under the discussion of equation (66).
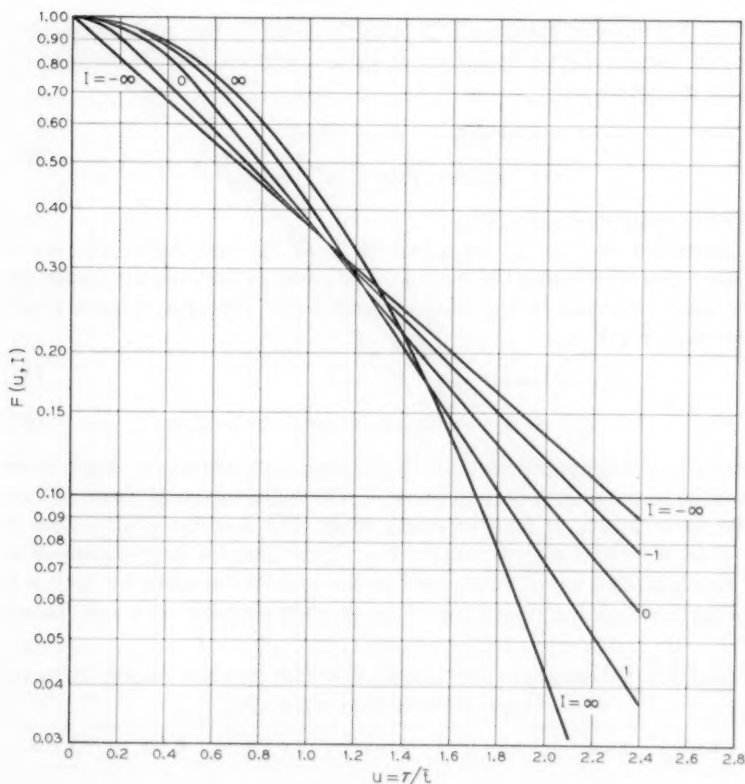
Fig. 1 — The probability $F(u, I)$ that $I(t) > I$ for an interval lasting longer than τ. $I(t)$ has a Gaussian power spectrum and $\bar{t}$ is given by equation (17).

D. S. Palmer[5] has obtained an approximate integral equation for the probability density $f(u, 0) = -\partial F(u, 0)/\partial u$ which has been improved by J. A. McFadden.[7] McFadden gives curves obtained by solving his equation numerically. They agree well with the experimental results of Ref. 6. For (19), $f(u, 0)$ appears to approach a value close to 1.25 as u approaches zero, whereas for (16), $f(u, 0)$ approaches zero.

The corresponding results for the envelope $R(t)$ may be stated as follows. Let $I(t)$ be a Gaussian noise current with the envelope $R(t)$ and the narrow-band power spectrum

$$w(f) = \sigma^{-1}(2\pi)^{-1/2} \exp\left[-(f - f_0)^2/2\sigma^2\right]. \qquad (20)$$

When $R(t)$ crosses downward through the level R, the probability that it will remain below R for more than τ seconds may be denoted by $F_r(u, R)$, where now

$$u = \tau/\bar{l}, \qquad \beta_r = (2\pi\sigma)^2,$$
$$\bar{l} = (R\sigma)^{-1}(2\pi)^{-1/2}(e^{R^2/2} - 1),$$
(21)

and $\bar{l}$ is computed from (10).

Estimates of $F_r(u, R)$ are given in Table III and the calculation of these values is discussed in Section VIII. Just as for $I(t)$, the values for the larger u's must be regarded as conjectures. The special cases corresponding to (18) are

$$F_r(u, R \to \infty) = e^{-u},$$
(22)

$$F_r(u, R \to 0) = (2/u)I_1(2/\pi u^2) \exp\,(-2/\pi u^2),$$
(23)

where $I_1(z)$ denotes a Bessel function of imaginary argument. Again these limiting forms do not depend on $w(f)$. The values given in Table III are plotted in Fig. 2. It may be shown from (23) that, for small values of u, $F_r(u, R \to 0)$ is approximately $1 - \frac{3}{16}\pi u^2$ and, for large values of u, it decreases like $2/\pi u^3$. Since the $\bar{l}$ in the u of (23) is given by (11), u is proportional to τ/R, and hence $F_r(u, R \to 0)$ depends on τ and R only

TABLE III — ESTIMATES OF $F_r(u, R)$ FOR THE NARROW-BAND, NORMAL-LAW POWER SPECTRUM (20)

u	$R = 0$	$R = 0.5$	$R = 1$	$R = 1.18$	$R = 2$	$R = \infty$
0	1.00	1.00	1.00	1.00	1.00	1.00
0.1	0.995	0.994	0.993	0.992	0.976	0.905
0.2	0.976	0.974	0.969	0.963	0.885	0.819
0.3	0.945	0.938	0.924	0.906	0.793	0.741
0.4	0.894	0.884	0.843	0.808	0.711	0.670
0.5	0.823	0.804	0.733	0.694	0.637	0.606
0.6	0.729	0.697	0.624	0.598	0.571	0.549
0.7	0.619	0.588	0.535	0.528	0.511	0.496
0.8	0.521	0.492	0.464	0.467	0.458	0.449
0.9	0.429	0.412	0.407	0.413	0.411	0.406
1.0	0.354	0.343	0.360	0.364	0.368	0.368
1.2	0.242	0.249	0.279	0.285	0.296	0.301
1.4	0.171	0.190	0.216	0.222	0.237	0.246
1.6	0.122	0.149	0.168	0.174	0.190	0.202
1.8	0.0890	0.116	0.130	0.135	0.153	0.165
2.0	0.0681	0.0899	0.101	0.106	0.123	0.135
2.2	0.0535	0.0701	0.0784	0.0826	0.0987	0.111
2.4	0.0425	0.0544	0.0608	0.0645	0.0792	0.0907
2.6	0.0340	0.0421	0.0472	0.0504	0.0636	0.0743
2.8	0.0269	0.0327	0.0366	0.0393	0.0510	0.0608
3.0	0.0220	0.0253	0.0284	0.0307	0.0410	0.0498
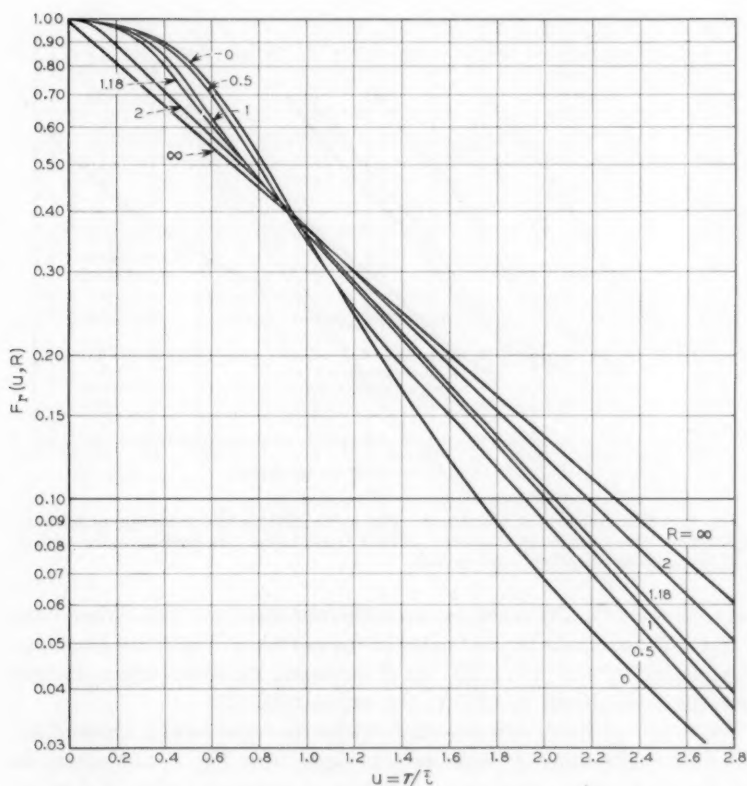
Fig. 2 — The probability $F_r(u, R)$ that $R(t) < R$ for an interval lasting longer than τ. $R(t)$ is the envelope of narrow-band Gaussian noise having the power spectrum (20).

through the ratio τ/R. The curves for $F_r(u, R \to 0)$ and $F_r(u, R \to \infty)$ cross near $u = 5.68$, where they both have the value 0.00340. This behavior is shown in Fig. 4.

IV. COMPARISON OF THEORY WITH EXPERIMENT

By combining the average fade lengths given by Table II and the fade length distributions $F_r(u, R)$ given by Table III, we obtain Fig. 3. In Fig. 3 the scales have been chosen so that the median value of $R(t)$ is 1.18 [see equations (15)], and so that fades below the median last one second, on the average. This leads to $\bar{t}_1 = 0.763$ seconds and $\beta_r = 4.53$

Fig. 3 — Distribution of the duration of fades. $R(t)$ is the envelope of narrow-band noise. The time scale is chosen so that fades below the median value of $R(t)$ have an average length of one second.

and makes $\sigma = 0.339$ in the associated power spectrum (20). When both R and τ/R are small, the ordinates of the curves on Fig. 3 are given approximately by $1 - 1.7(\tau/R)^2$. As R increases, the distribution changes from the form shown by (23) to the exponential (22).

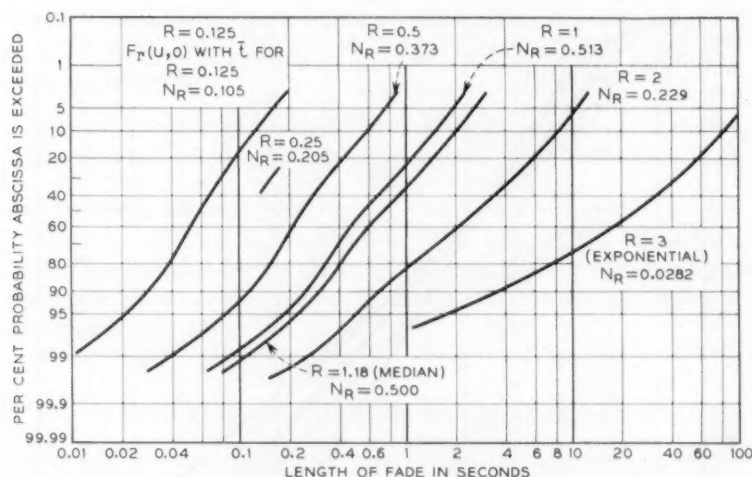When the available experimental results for rapid fading are studied, it is found that some of them seem to agree with Fig. 3, but others do not. It is not hard to find reasons for lack of agreement when it does occur. One is that we have arbitrarily taken the normal-law power spectrum (20) for our calculations. This gives fade length distributions $F_r(u, R)$ of the sort shown in Fig. 2. These distributions appear to be typical of those obtained for $w(f)$'s which (i) decrease rapidly for frequencies outside the nominal band and (ii) do not become excessively large for frequencies near f_0 (such frequencies correspond to slow fluctuations in the envelope). For $w(f)$'s which do not satisfy these requirements, say one which behaves like

$$w(f) = \frac{2k}{k^2 + 4\pi^2(f - f_0)^2} \tag{24}$$

when f is near f_0, k being very small, the work in Appendix IV suggests that the distribution curves have the quite different shape shown in Fig. 13. When the experimental curves differ from those of Fig. 2, they tend to look like those of Fig. 13.
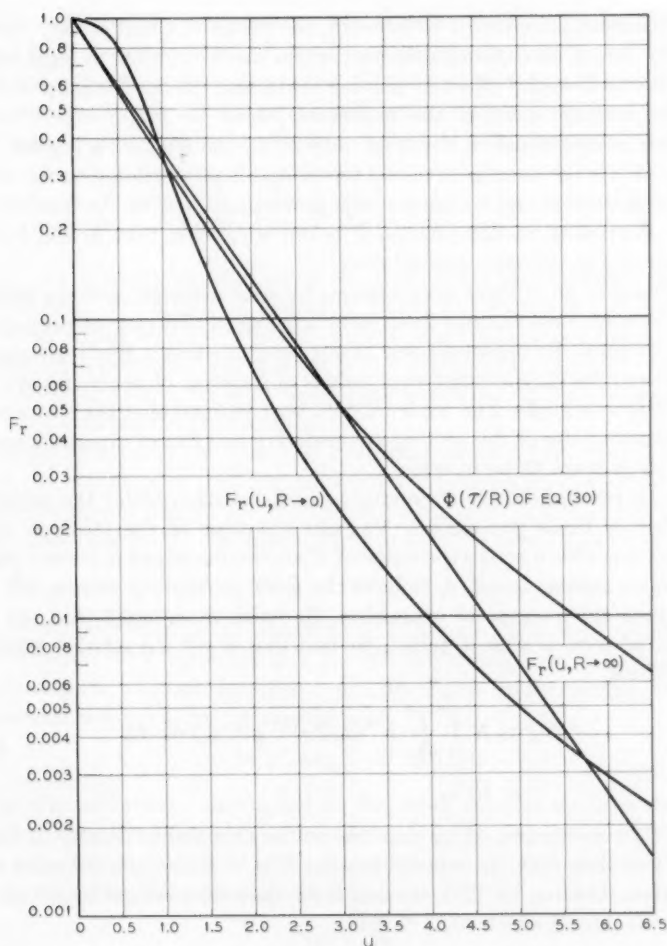
Fig. 4 — Plots of $F_r(u, R)$ for special cases. $F_r(u, R \to \infty)$ is given by (22). $F_r(u, R \to 0)$ corresponds to fixed values of the median and fading rate, and is given by (23). $\Phi(\tau/R)$ of (30) corresponds to fixed median but variable fading rate.

A second reason for the lack of agreement is that Fig. 3 refers to fixed values of the parameters b and β_r where $b^{1/2}$ is a measure of the median value $(R_m = 1.18b^{1/2})$ and $(\beta_r/b)^{1/2}$ is a measure of the rate of fluctuation [the average number of times $R(t)$ passes down across the median value in one second is $N_{R_m} = 0.235(\beta_r/b)^{1/2}$]. Actually, assuming

the Gaussian noise model to be valid, the values of b and β_r may change slowly during an experimental run. When the fluctuation is regarded as due to the Doppler effect of moving scatterers, $(\beta_r/b)^{1/2}$ is proportional to the average speed of the scatterers, which we might expect to be roughly proportional to the wind velocity in the scattering region. The value of b is the average received power and is proportional to the number of scatterers and to the average power scattered to the receiver by each. According to this picture, it is not surprising that β_r and b may change during an experimental run.

In spite of all of these uncertainties we shall attempt (perhaps rashly) to draw some conclusions from our formulas in the case of deep fades (R near zero). In this case a few of our formulas take a relatively simple form, and we may average over values of b and β_r or otherwise use the formulas as a guide. The reader should bear in mind that the statements in the remainder of this section are tentative and future experimentation may show them to be in error.

As an example of the averaging process, let us consider the expected number of times per second, N_R, the envelope of the received wave crosses downward across the value of R microvolts where R is very small. When we assume b and β_r to have the joint probability density $g(b, \beta_r)$ and take the average of expression (9) (with R so small that the exponential term is almost unity), we find that if $g(b, \beta_r)$ satisfies suitable conditions,

$$N_R \approx R \int_0^\infty \int_0^\infty b^{-1}(\beta_r/2\pi)^{1/2} \, g(b, \beta_r) \, db \, d\beta_r \qquad (25)$$

$$= RC_1 .$$

Here C_1 is a constant which depends on the probability density of b and β_r. By dealing with the probability that $R(t) < R$ in much the same way and then dividing by (25), we can show that the average length of the intervals during which $R(t) < R$ is

$$\bar{l} \approx RC_2 \qquad (26)$$

where C_2 is a constant of the same nature as C_1 and R is again small.

Equation (25) says that, for a given experimental run, the average number of times per minute the signal drops below $R/2$ microvolts should be half the number of times it drops below R microvolts, R being very small. Furthermore, from (26), fades below $R/2$ microvolts should last, on the average, only half as long as the fades below R. These relations should hold for small values of R even though the experimentally observed ratios for the larger values of R may not agree with Table II.

The same sort of argument may be applied to the fade length distribution. As R becomes small, $F_r(u, R)$ approaches a definite function of $u = \tau/\bar{t}$, say $\varphi(u)$, provided u is not too large. Let R have the same significance as in (25). Then the probability that $R(t)$ will remain less than R for an interval lasting longer than τ seconds is approximately

$$\frac{R}{N_R} \int_0^\infty \int_0^\infty b^{-1} \sqrt{\frac{\beta_r}{2\pi}} \, \varphi\left[\frac{2\tau}{R} \sqrt{\frac{\beta_r}{2\pi}}\right] g(b, \beta_r) \, db \, d\beta_r = \Phi(\tau/R), \quad (27)$$

where N_R is given by (25). Thus the probability depends on τ and R only through the ratio τ/R when R is small. The curves of Fig. 3 corresponding to the smaller values of R show this behavior (as indeed they should). For example, the curve for $R = 0.125$ is obtained almost exactly by shifting the curve for $R = 0.50$ to the left by an amount corresponding to quartering the fade lengths.

As an illustration of how the averaging process (27) may change $F_r(u, R)$, we suppose that the median level remains fixed at $R = 1.18$ while the rate of fluctuation changes in such a way as to formally carry the normal-law power spectrum (20) over into the power spectrum.

$$w(f) = \pi k^2 \left[k^2 + 4\pi^2(f - f_0)^2\right]^{-3/2}.$$

This particular example is chosen so as to simplify the steps which follow. Instead of introducing $g(b, \beta_r)$ and averaging over β_r it is more convenient to average over the σ of (20). Some mathematical experimentation indicates that we should take

$$(k^2/4\pi^2\sigma^3) \exp \left[-k^2/8\pi^2\sigma^2\right] d\sigma \quad (28)$$

to be the probability that σ lies in the small interval $d\sigma$ because this choice makes the integral from $\sigma = 0$ to $\sigma = \infty$ of the product of (28) and (20) give the required $w(f)$. When we give β_r the value $(2\pi\sigma)^2$ indicated by (21), the φ function in (27) becomes

$$\varphi\left[\frac{\tau\sigma 2\sqrt{2\pi}}{R}\right] \quad (29)$$

where we take $\varphi(u)$ to be the limiting form of $F_r(u, R \to 0)$ given by (23). The integral (25) now gives $N_R = Rk/2$ and the integral in (27) becomes the integral from $\sigma = 0$ to $\sigma = \infty$ of the product of (28) and (29). The integration can be performed and gives

$$\Phi(\tau/R) = \left[1 + \frac{u^2}{4}\right]^{-1/2} \left[1 + \frac{u^2}{2} + u \sqrt{1 + \frac{u^2}{4}}\right]^{-1}, \quad (30)$$

where $u = k\tau/R$. This is the fade length distribution, when R is small,

which results when the median level remains fixed but the rate of fading fluctuates according to the law (28). The average value $\bar{l}$ of the fade length works out to be R/k, so that $u = \tau/\bar{l}$ in (30), in agreement with our earlier notation.

If we were to take σ^ν in place of σ^3 in (28) (and change the multiplying constant) we could obtain a more general, but also more complicated, expression for $\Phi(\tau/R)$. When $\nu = 2$ the power spectrum (24) is obtained but the integrals (25) and (27) corresponding to N_R and $\Phi(\tau/R)$ do not converge.

The expression (30) is plotted as a function of u in Fig. 4, along with $F_r(u, R \rightarrow \infty)$ and $F_r(u, R \rightarrow 0)$, given respectively by (22) and (23), for comparison. As u increases, (30) ultimately decreases as $2/u^3$. It is seen that the $\Phi(\tau/R)$ given by (30) and $F_r(u, R \rightarrow 0)$ [from which (30) is derived] are quite different in appearance for small values of u. The distribution for fixed median and fixed fading rate is $F_r(u, R \rightarrow 0)$, while the $\Phi(\tau/R)$ of (30) corresponds to a fixed median but a variable fading rate. It appears from (27) that many different $\Phi(\tau/R)$'s [of which (30) is only one] may be obtained by averaging over the median values and fading rates with different weighting functions $g(b, \beta_r)$. At this stage we are unable to tell from either theory or experiment whether a different $\Phi(\tau/R)$ will be required for each set of data, or whether there are only two or three universal types of $\Phi(\tau/R)$'s which will fit, for small R, all kinds of fading data.

Suppose that we are given a set of experimental results on short-term fading. How can we go about analyzing them to see whether they have anything in common with our theory?

One of the easiest things to test is how N_R, the average number of fades per second below the level R, depends on R. According to (25), N_R should be proportional to R when R is small, and, if the short-term median value and the rate of fading have remained fixed throughout the experiment, the ratios of N_R/N_1 should agree with those of Table II. From (26) the same is true of $\bar{l}$ and $\bar{l}/t_1$. However, when the short-term median value and the rate of fading fluctuate, N_R and $\bar{l}$ should still be proportional to R when R is small, but the ratios in Table II may no longer hold. It should be noted that Table II lists values of both N_R/N_1 and $\bar{l}/\bar{l}_1$, and it may turn out that one set of ratios will agree better with experiment than the other. This may happen because N_R and $\bar{l}$ do not depend on b and β_r in the same way. N_R depends on both β_r and b, $\bar{l}$ only on β_r, and the probability that $R(t) < R$ (which, for small R, becomes $R^2/2b$) depends only on b. In fact, the deviations from Table II and from the Rayleigh distribution should give clues to the variation of b and β_r.

When experimental fade-length distributions for several levels of R are available, it is instructive to plot them on semi-log paper as a function of τ/R in much the same way as $F_r(u, R)$ is plotted as a function of u in Fig. 2. As R becomes small the curves should tend to coincide, at least when τ/R is not too large. This follows from (27). If the limiting curve can be made to coincide with the $R = 0$ curve in Fig. 2 by changing the scale of τ/R, then it is likely that the other curves in Fig. 2 and the curves of Fig. 3 will be valid. This indicates that b and β_r did not change much during the experiment.

If, on the other hand, the limiting curve looks like those of Fig. 13 or like the $\Phi(\tau/R)$ of Fig. 4, the probability of long fades is much greater. In this case one should plot on semi-log paper R times the probability that the fade length exceeds τ versus τ/R^2. For large τ and small R, the points should tend to lie on a common curve, which becomes a straight line as τ/T^2 becomes large. This behavior is suggested by one of the conjectures of Appendix IV concerning long fades when $w(f)$ behaves like (24) near f_0, k being very small (i.e., much power in the frequencies corresponding to slow fluctuations). It is assumed that $w(f)$ is still such that $\bar{l}$ exists. For $w(f)$'s of this sort the conjecture says that when τ is very large and R is small

$$P(\tau, R) = F_r(u, R) \rightarrow 4\bar{l}kR^{-2} \sum_{i=1}^{\infty} e^{-\alpha_i k \tau} \qquad (31)$$

where $P(\tau, R)$ is the probability that a fade below level R will last longer than τ seconds. In (31) the median value of $R(t)$ is 1.18 and α_i is $j_i^2 R^{-2}$, j_i being the ith zero of $J_0(x)$.

From (11), $\bar{l}$ is proportional to R and it follows that $RP(\tau, R)$ tends to be a function of τ/R^2. This is the basis for the method of plotting described above. When τ is very large, only the first term in (31) is important, and it gives a straight line on the semi-log paper. As R increases, α_1 is no longer given by $5.78 R^{-2}$. For $R = 1$, $\alpha_1 \approx 5.0$ and for $R = $ median $= 1.18$, $\alpha_1 \approx 3.25$.

Unfortunately we do not have any clear-cut rules to tell when, as τ increases, the fade-length distribution changes from a function of τ/R to $1/R$ times a function of τ/R^2.

Sometimes experimental results are given in the form of curves showing the average number of fades per hour, say $H(\tau, R)$, below the level R which last longer than τ seconds. According to (25) and (27), $H(\tau, R)$ should take the form $3600 N_R \Phi(\tau/R) = CR\Phi(\tau/R)$, where C is a constant, R is small, and τ/R not too large. Under the conditions leading to (31), we expect $H(\tau, R)$ to be of the form $\Psi(\tau/R^2)$ when τ is large and R is small. In fact, if we assume (31) and note that $N_R \bar{l}$ approaches

$R^2/2$ [b is one in (31), so that the median value here is 1.18],

$$H(\tau, R) \rightarrow 7200 \, k \, \exp\left[-5.78 \, k\tau/R^2\right]. \tag{32}$$

Since this expression contains only one parameter, k, and is based on a number of assumptions, it is probably too simple to agree well with experiment. Nevertheless, if the experimental data for $H(\tau, R)$ passes the test of being of the form $\Psi(\tau/R^2)$ for large τ and small R, it would be interesting to go one step farther and see how well the data can be fitted by a suitable choice of k in (32).

It is interesting to examine Fig. 13 of Ref. 1 in the light of the above results. This figure, with some dashed lines added, is reproduced here as Fig. 5. The three solid lines give the observed (for 505 mc) number of fades per hour which were more than 5, 10 and 15 db below the hourly median level and lasted longer than the time intervals shown on the abscissa scale. The two dashed straight lines labeled (a) show the 5- and 15-db curves obtained by applying our results to the 10-db curve in the following way. The number of fades per hour of depth R of duration longer than τ seconds is $H(\tau, R)$ which, when R is small and τ/R not too large, may be written as

$$H(\tau, R) = CR\Phi(\tau/R), \tag{33}$$

where C is a constant. The values of R corresponding to the 5-, 10-, 15-db curves are in the ratio of 1.78, 1, 0.56. Since the curves of Fig. 5 are plots of $H(\tau, R)$ versus τ, (33) gives a relation between them. Thus, suppose the coordinates (τ_1, H_1) of a point on the R_1 curve are known. Then the coordinates of the point (τ_2, H_2) corresponding to the same value of τ/R on the R_2 curve are given by

$$\begin{aligned} \tau_2 &= \tau_1 R_2/R_1, \\ H_2 &= H_1 R_2/R_1. \end{aligned} \tag{34}$$

The 15-db dashed line (a) was obtained by taking the solid 10-db line to be the R_1 curve and using (34) with $R_2/R_1 = 0.56$. The 5-db dashed line (a) was obtained in a similar way from the 10-db line by taking $R_2/R_1 = 1.78$, even though this may take us into the region where our approximation begins to fail. The dashed lines (a) are drawn only for the smaller values of τ/R because of our assumption that (33) may not hold if τ/R is too large.

Under certain conditions we may expect $H(\tau, R)$ to depend only on the ratio τ/R^2 when τ is large and R is small. The dashed straight lines labeled (b) are obtained from the 10-db curve by assuming $H(\tau, R)$ to
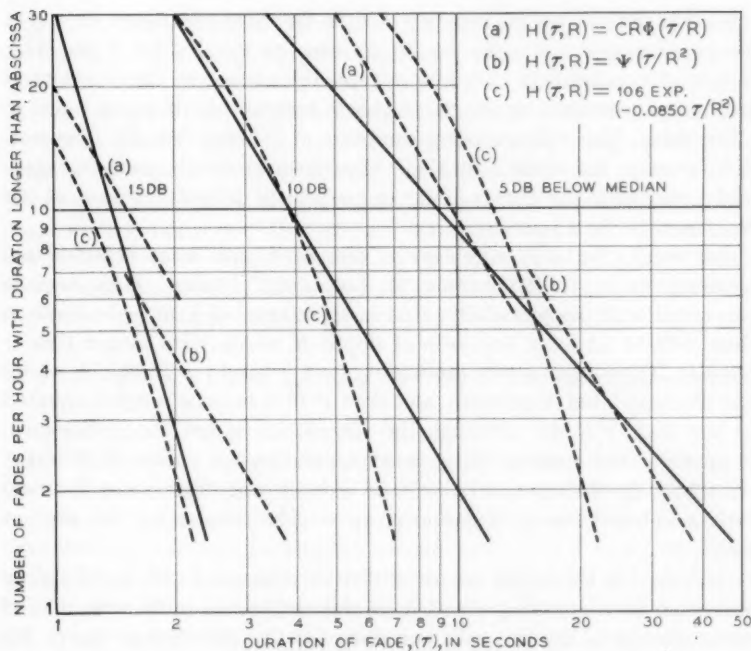
Fig. 5 — Number vs. duration of fast fades at 505 mc.

be of the form $\psi(\tau/R^2)$ and proceeding in much the same way as in the calculation of the lines (a). The lines (b) are drawn only for the larger values of τ/R^2 because τ is supposed to be large.

The curved dashed lines labeled (c) in Fig. 5 are plots of $H(\tau, R)$ computed from (32), with k chosen so that $H(\tau, R)$ agrees with the experimental 10-db result at the ordinate corresponding to ten fades per hour. Since the value of R, 10 db below the median value of 1.18, is 0.372, and since $\tau = 3.8$ seconds corresponds to ten fades per hour, the equation to determine k is

$$H(3.8, 0.372) = 10. \tag{35}$$

This yields two values of k, namely 0.0018 and 0.0147. The value 0.0147 was used in plotting the curves (c) in Fig. 5, since the other value gives curves with much smaller slopes.

The difference between the dashed curves in Fig. 5 and the corresponding solid curves is noticeable. An examination of the original data indi-

cates that some of the discrepancy may be due to the fact that, in one of the experimental runs, N_R for 10 db exceeds the N_R for 5 db. This completely contradicts (25) and is probably caused by the short-term median (as measured by the parameter b) temporarily dropping down to a low value. This violates our assumption of R being "small". The relation between the 10-db and 15-db experimental results seems to agree fairly well with the theory, but comparison is difficult because of the scantiness of the 15-db data.

Our results for $R(t)$ may also be compared with some experimental results given by R. R. Favreau, H. Low and I. Pfeffer.[6] Their work is concerned with the probability that in an interval of T seconds duration there will be at least one fade of depth R which lasts longer than τ seconds. They express this probability as $1 - \exp\,[-\,P'(0)f_bT]$, where f_b is the associated bandwidth, and their $P'(0)f_b$ may be roughly equated to our $N_RF_r(\tau/\bar{t},\,R)$. Although the comparison is not straightforward, it appears that there is fairly good agreement for values of R larger than 0.25. As R decreases below 0.25 a disturbing discrepancy between experiment and theory seems to creep in. The reasons for this are not clear.

Although in this paper we are primarily concerned with rapid fading and have been speaking about it in this section up until now, it is of some interest to try to apply our theory to the slow fades, i.e., to the distribution of the hourly medians. Since the hourly medians have a distribution which is nearly normal in decibels, it is natural to compare the observed statistics with those of a Gaussian noise current $I(t)$. In this comparison "db above the median" plays the same role as "amperes".

When the fade-length distributions given in Figs. 6 and 7 of Ref. 1 are plotted on semi-log paper the resulting points tend to lie on straight lines instead of on curves shaped like those of Fig. 1. This exponential distribution of fade lengths indicates that the power spectrum of $I(t)$ is more like the ones discussed in Section IX and Appendix IV than like the normal type shown in equation (16). A stochastic process which does have an exponential distribution of fade lengths may be constructed by drawing a value at random from a normal universe every hour on the hour. However, the successive values in such a process would be uncorrelated. This is not true for the hourly medians.

Even though we have little justification for doing so, it is interesting to compare the separation between the experimental curves for various depths of fades with the separation predicted by the analogue of equa-

tion (27). The analogue states that the probability of $I(t)$ remaining less than $-I$ for an interval lasting longer than τ is approximately of the form $\Phi(\tau I)$ when I is large. According to this rule, the -10-db curves in Figs. 6 and 7 of Ref. 1 may be obtained by shifting the -5-db curves to the left by an amount corresponding to changing the time scale in the ratio 5 to 10 (or 1 to 2). Actually, a ratio of 1 to 3 would be required to agree with the experimental results. The analogue of (31) suggests in a vague way that when τ becomes large, I remaining large, the function $\Phi(\tau I)$ might change into a function of τI^2. This would correspond to changing the time scale in the ratio of 1 to 4. However, not much significance can be attached to these figures.

V. DISTRIBUTION OF THE LENGTHS OF THE INTERVALS DURING WHICH $I(t) > I$

In Section II we have given expressions for the average length $\bar{l}$ and the expected number per second N_I of the intervals during which $I(t)$ exceeds the value I. As a first step in finding the probability $p(\tau, I) \, d\tau$ that the length of such an interval lies between τ and $\tau + d\tau$, we obtain an approximation, $p_1(\tau, I) \, d\tau$, which agrees well with $p(\tau, I) \, d\tau$ for small values of τ. The expression $p_1(\tau, I) \, d\tau$ is the probability that, if $I(t)$ passes upward through the value I at time t_1, it will pass downward through I in the infinitesimal interval t_2, $t_2 + d\tau$ where $t_2 = t_1 + \tau$.

An expression for $p_1(\tau, I)$ may be obtained by the method outlined in Section 3.4 of Ref. 2. Thus, the probability that $I(t)$ will pass upwards through the level I in t_1, $t_1 + dt_1$ and downwards through I in t_2, $t_2 + dt_2$ is

$$- dt_1 \, dt_2 \int_0^\infty dI_1' \int_{-\infty}^0 dI_2' I_1' I_2' p(I, I_1', I_2', I), \qquad (36)$$

where $p(I_1, I_1', I_2', I_2)$ denotes the probability density of $I(t_1)$, and $I'(t_1)$, $I'(t_2)$, $I(t_2)$ and the primes denote time derivatives. Then

$$p_1(\tau, I) \, dt_2 = \frac{\text{expression (36)}}{\text{prob of } I(t) \text{ passing upward through } I \text{ in } t_1, \, t_1 + dt_1} \qquad (37)$$

and we obtain

$$p_1(\tau, I) = -2\pi\beta^{-1/2} e^{I^2/2} \int_0^\infty dI_1' \int_{-\infty}^0 dI_2' I_1' I_2' p(I, I_1', I_2', I). \qquad (38)$$

Here β is given by equation (4) and we have taken $b = \overline{I^2(t)} = 1$.

The probability density $p(I_1, I_1', I_2', I_2)$ is

$$
(2\pi)^{-2} M^{-1/2} \exp \Big\{ -\frac{1}{2M} [M_{11} I_1^2 + 2M_{12} I_1 I_1' + 2M_{13} I_1 I_2'
$$
$$
+ 2M_{14} I_1 I_2 + M_{22} I_1'^2 + 2M_{23} I_1' I_2' \tag{39}
$$
$$
+ 2M_{24} I_1' I_2 + M_{33} I_2'^2 + 2M_{34} I_2' I_2 + M_{44} I_2^2] \Big\},
$$

where M_{ij} is the cofactor of the ijth element in the determinant

$$
M = \begin{vmatrix} 1 & 0 & m' & m \\ 0 & \beta & -m'' & -m' \\ m' & -m'' & \beta & 0 \\ m & -m' & 0 & 1 \end{vmatrix} \tag{40}
$$

in which we have simplified the notation of Ref. 2 by writing

$$
m = \psi_\tau = \int_0^\infty w(f) \cos 2\pi f\tau \, df, \tag{41}
$$
$$
m' = dm/d\tau, \qquad m'' = d^2 m/d\tau^2, \qquad \beta = [-m'']_{\tau=0}.
$$

The values of the M_{ij}'s and some relations among them are listed in Appendix I. When I_1 and I_2 are set equal to I and use is made of the fact that some of the M_{ij}'s are equal, the exponent in (39) reduces to $-1/(2M)$ times

$$
2(M_{11} + M_{14}) I^2 + 2(M_{12} + M_{24})(II_1' - II_2') + M_{22}(I_1'^2 + I_2'^2)
$$
$$
+ 2M_{23} I_1' I_2'. \tag{42}
$$

Replacing I_1', I_2' in (42) by $x_1 - a_1$, $x_2 - a_2$, choosing the a's so as to make the first order terms in x_1 and x_2 vanish and using

$$
(M_{12} + M_{24})(1 + m) = m'(M_{22} - M_{23})
$$
$$
M = (M_{11} + M_{14})(1 + m) - m'(M_{12} + M_{24}) \tag{43}
$$

gives

$$
a_1 = -a_2 = m' I/(1 + m) \tag{44}
$$

and converts (42) into

$$M_{22}(x_1^2 + x_2^2) + 2M_{23}x_1x_2 + 2I^2M(1 + m)^{-1}. \tag{45}$$

When the change of variable corresponding to

$$x(1 - r^2)^{-1/2} = x_1(M_{22}/M)^{1/2}$$

$$y(1 - r^2)^{-1/2} = -x_2(M_{22}/M)^{1/2} \tag{46}$$

$$r = M_{23}/M_{22} = \frac{m''(1 - m^2) + mm'^2}{\beta(1 - m^2) - m'^2}$$

is made in (38) it goes into

$$p_1(\tau, I) = M_{22}\beta^{-1/2}(1 - m^2)^{-3/2} \exp\left[\frac{I^2}{2} - \frac{I^2}{1 + m}\right] J(r,h), \tag{47}$$

where

$$J(r, h) = \frac{1}{2\pi s} \int_h^\infty dx \int_h^\infty dy(x - h)(y - h)e^z,$$

$$h = \frac{m'I}{1 + m}\left[\frac{1 - m^2}{M_{22}}\right]^{1/2}, \qquad s = (1 - r^2)^{1/2}, \tag{48}$$

$$M_{22} = \beta(1 - m^2) - m'^2,$$

$$z = -\frac{x^2 + y^2 - 2rxy}{2(1 - r^2)} = -2^{-1}s^{-2}(x^2 + y^2 - 2rxy), \tag{49}$$

and we have used

$$(1 - m^2)M = M_{22}^2 - M_{23}^2 = M_{22}^2(1 - r^2). \tag{50}$$

$J(r, h)$ may be expressed in terms of the tabulated double integral

$$K(r, h) = \frac{1}{2\pi s} \int_h^\infty dx \int_h^\infty dy\, e^z \tag{51}$$

as follows. We have

$$\frac{\partial}{\partial x} e^z = s^{-2}(ry - x)e^z,$$

$$\frac{\partial}{\partial y} e^z = s^{-2}(rx - y)e^z.$$

Elimination of the y terms on the right leads to

$$\left(\frac{\partial}{\partial x} + r\frac{\partial}{\partial y}\right)e^z = -xe^z. \tag{52}$$

In much the same way, elimination of the x^2 and y^2 from the second partial derivatives gives

$$\left[r \frac{\partial^2}{\partial x^2} + r \frac{\partial^2}{\partial y^2} + (1 + r^2) \frac{\partial^2}{\partial x \partial y} \right] e^z = xye^z - re^z. \tag{53}$$

When the integrand in $J(r, h)$ is written as $(xy - hx - hy + h^2)e^z$ and use is made of (52), (53) and the symmetry in x and y, we obtain

$$J(r, h) = \frac{1}{2\pi s} \int_h^\infty dx \int_h^\infty dy \left[2r \frac{\partial^2}{\partial y^2} + (1 + r^2) \frac{\partial^2}{\partial x \partial y} \right.$$

$$\left. + r + 2h(1 + r) \frac{\partial}{\partial y} + h^2 \right] e^z \tag{54}$$

$$= (r + h^2)K(r, h) + (s/2\pi) \exp\left[-\frac{h^2}{1 + r} \right] - h(2\pi)^{-1/2} e^{-h^2/2}[1 - P(a)]$$

where $P(a)$ is the error integral defined by (3) and

$$a = h \left[\frac{1 - r}{1 + r} \right]^{1/2}, \qquad \frac{h^2}{1 + r} = \frac{a^2 + h^2}{2}. \tag{55}$$

Values of $K(r, h)$ may be obtained from tables[8] of d/N. This is true because d/N is the function of the three variables r, h, k obtained from the right-hand side of (51) by changing the lower limit of integration in the y integral from h to k. The tables do not extend to negative values of h, and $K(r, h)$ for these values must be obtained from

$$K(r, -h) = K(r, h) + P(h). \tag{56}$$

From (56) it may be shown that

$$p_1(\tau, -I) = p_1(\tau, I) + \frac{M_{22}}{\beta^{1/2}(1 - m^2)^{3/2}} \left[(r + h^2)P(h) + \frac{2he^{-h^2/2}}{\sqrt{2\pi}} \right] \tag{57}$$

$$\exp\left[\frac{I^2}{2} - \frac{I^2}{1 + m} \right]$$

where h depends on I in accordance with (48).

The following special values of $K(r, h)$ and $J(r, h)$ may be derived directly from known results.

1. When $h = 0$

$$K(r, 0) = \frac{1}{2} - \frac{\cos^{-1} r}{2\pi}, \qquad 0 < \cos^{-1} r \leqq \pi \tag{58}$$

$$J(r, 0) = rK(r, 0) + \frac{s}{2\pi}.$$

2. When $r = 0$

$$s = 1, \qquad a = h,$$

$$K(0, h) = \tfrac{1}{4}[1 - P(h)]^2,$$

$$J(0, h) = \left\{ \frac{h}{2}[1 - P(h)] - (2\pi)^{-1/2}e^{-h^2/2} \right\}^2.$$

(59)

3. When r approaches unity from below we have, in the limit,

$$s = 0, \qquad a = 0,$$

$$K(1, h) = \tfrac{1}{2}[1 - P(h)],$$

$$J(1, h) = 2^{-1}(1 + h^2)[1 - P(h)] - h(2\pi)^{-1/2}e^{-h^2/2}.$$

(60)

4. When h approaches $-\infty$, $K(r, \infty)$ approaches one and $J(r, h)$ is approximately $h^2 + r$.

As τ becomes very large $p_1(\tau, I)$ approaches the value of N_I which is

$$(2\pi)^{-1}\beta^{1/2} \exp(-I^2/2).$$

(61)

This follows from considerations of the type which led to (5). It may also be derived from (47), (48) and the fact that m, m', m'' all approach zero as τ becomes large. In this case M_{22} approaches β, and r and h become zero.

When I is zero, equation (47) reduces to a result given in Ref. 2, namely

$$p_1(\tau, 0) = M_{22}\beta^{-1/2}(1 - m^2)^{-3/2}[r(\pi - \cos^{-1} r) + s]/2\pi,$$

(62)

where $0 < \cos^{-1} r < \pi$.

Expressions for the various parameters entering the expression for $p_1(\tau, I)$ when τ is small are given in Appendix II. Combining the results of Appendix II with (47), we obtain

$$p_1(\tau, I) \approx (\tau B/4\beta) \exp(-I^2\tau^2\beta/8)J(1, h)$$

$$h \approx -I\beta B^{-1/2}$$

(63)

where τ is assumed to be small and B is given by one of equations (125). There are two cases to be considered. In the first one, I remains fixed as τ approaches zero. It then follows that $p_1(\tau, I)$ approaches zero linearly with τ according to

$$p_1(\tau, I) \approx (\tau B/4\beta)J(1, h)$$

(64)

where $J(1, h)$ and h are given by (60) and (63). In the second case, I is positive and increases as τ decreases so as to keep the order of τI at

unity. Then h approaches $-\infty$ and $J(1, h)$ is approximately h^2, with the result that

$$p_1(\tau, I) \approx (\beta\tau I^2/4) \exp(-I^2\tau^2\beta/8). \qquad (65)$$

This expression may also be obtained from (47) by first letting I become large (assuming m' to be negative so that h becomes large and negative) and then letting τ become small.

When the right-hand side of (65) is integrated from $\tau = 0$ to $\tau = \infty$ the result is unity. Since the same is true of the exact probability density $p(\tau, I)$ [to which $p_1(\tau, I)$ is a first approximation], and since $p_1(\tau, I) \geqq p(\tau, I)$, we are led to expect that

$$p(\tau, I) \to (\beta\tau I^2/4) \exp(-I^2\tau^2\beta/8) \qquad (66)$$

when I becomes large and positive (and τI remains fixed). A formal proof of (66) has been constructed on the assumptions (i) that $p(\tau, I)$ ultimately decreases exponentially as τ becomes large, and (ii) that m' remains negative for all positive values of τ. The second assumption seems to be required by the method of proof rather than by nature. We shall not give the proof here because of its length. Instead, we shall outline a method by which (66) may be obtained directly.* This method is due to D. S. Palmer and grew out of correspondence with him.

When I is very large, the intervals during which $I(t) > I$ are very short. The length of an interval starting at $t = t_1$ is, to a second degree Taylor approximation,

$$l = 2I'(t_1)/[-I''(t_1)].$$

The joint probability density of $I(t)$ and its first two derivatives is

$$(2\pi)^{-3/2}(\beta B)^{-1/2} \exp\left[-\frac{I^2}{2} - \frac{I'^2}{2\beta} - \frac{(I'' + \beta I)^2}{2B}\right]$$

where B is given by (125). Since I'' is a random variable whose mean value $-\beta I$ is large compared to its standard deviation, we may put $l = 2I'(t_1)/\beta I$.

The chance that $I(t)$ will pass upwards through the value I in the interval t_1, $t_1 + dt$ with a slope between I' and $I' + dI'$ is equal to the chance that, at time t_1, $I(t)$ lies between I and $I - I'dt$ and has a slope in I', $I' + dI'$. This chance is

$$\frac{(I' \, dt)(dI')}{2\pi\sqrt{\beta}} \exp\left[-\frac{I^2}{2} - \frac{I'^2}{2\beta}\right].$$

---

* Another proof has subsequently been devised by Mark Kac and D. Slepian. Also see the end of Appendix III.

Expressing I' in terms of l shows that the chance of $I(t)$ passing up through I in t_1, $t_1 + dt$ and starting an $I(t) > I$ interval whose length lies between l and $l + dl$ is

$$[\beta^{3/2}I^2l \; dl \; dt/8\pi] \exp\left[-\frac{I^2}{2} - \frac{\beta I^2 l^2}{8}\right]. \qquad (67)$$

Our goal of obtaining the probability density (65) for the lengths of the $I(t) > I$ intervals, I very large, is now within reach. All that remains to be done is to divide (67) by the chance $[\beta^{1/2}dt/2\pi] \exp[-I^2/2]$ that $I(t)$ will pass upwards across I in t_1, $t_1 + dt$.

In much of the following work it is convenient to use the variable u, defined by

$$u = \tau/\bar{l}, \qquad \bar{l} = \pi\beta^{-1/2}[1 - P(I)] \exp(I^2/2), \qquad (68)$$

where the expression for the average duration $\bar{l}$ is obtained by setting $b = 1$ in (6). The probability density $f(u, I)$ of u is related to the probability density $p(\tau, I)$ of τ by

$$f(u, I) = \bar{l}p(u\bar{l}, I), \qquad (69)$$

and the cumulative probability

$$F(u, I) = \int_u^\infty f(u, I) \; du \qquad (70)$$

has already appeared, for the special $w(f)$ given by (16), in Section III. The first approximation to $f(u, I)$ corresponding to $p_1(\tau, I)$ is

$$f_1(u, I) = \bar{l}p_1(u\bar{l}, I). \qquad (71)$$

When u becomes very large $f_1(u, I)$ approaches the valuê $[1 - P(I)]/2$.

The function $f(u, I)$ has the following limiting forms:

$$f(u, I) \to (u\bar{l}^2B/4\beta)J(1, h), \qquad u \to 0$$

$$f(u, I) \to (\pi u/2) \exp(-u^2\pi/4), \qquad I \to \infty \qquad (72)$$

$$f(u, I) \to \exp(-u), \qquad I \to -\infty, u > 0.$$

The first of these follows from (64), the second from (66) and the second of equations (6) and the third expresses the exponential behavior which sets in after $f(u, I)$ has passed through its maximum value (which occurs close to $u = 0$ when I is large and negative). The exponential law is followed because most of the intervals are so long that the probability of an interval of length τ ending in τ, $\tau + d\tau$ is independent of τ.

VI. APPROXIMATIONS FOR LONG INTERVALS

It is generally assumed that, for large values of τ, $p(\tau, T)$ decreases exponentially. This assumption is supported by the following informal argument, which is much the same as the one used at the end of Section V for the case $I = -\infty$. When $I(t)$ remains greater than I for an interval long in comparison with the effective range of the autocorrelation, $I(t)$ "forgets" when it entered the region $I(t) > I$. After this amnesia occurs it wanders about in the region in a fashion independent of its entrance time, until it happens to drop below the value I and thus end the interval. The exponential decrease is obtained when it is assumed that the chance of dropping below the value I in the interval $t, t + \Delta t$ depends only on Δt and not on how long $I(t)$ has been in the region.

However, no one seems to have given an exact and usable asymptotic formula for this exponential decrease, although progress in this direction has been made by Kuznetsov, Stratonovich and Tikhonov.[9] When applied to our problem, their result gives $F(u, I)$ as an exponential function of an infinite series whose nth term is a multiple integral of order $n - 1$. When τ becomes large each term approaches the form $a_n \tau + b_n$, where a_n and b_n are independent of τ. Here we shall be content to connect, by eye, the curve showing our estimates of $p(\tau, I)$ for small values of τ (obtained from $p_1(\tau, I)$) with an exponential curve chosen so as to (i) make the area under $p(\tau, I)$ unity and (ii) give the correct average value for τ. It is hoped that this will give an estimate of some use in engineering applications, even though it may not contribute a great deal to our understanding of the behavior of $p(\tau, I)$ for very large values of τ.

We shall work with $u = \tau/\bar{l}$, its probability density $f(u, I)$ defined by (69), and the approximation $f_1(u, I)$ defined by (71). The procedure consists of calculating

$$F_1(u, I) = 1 - \int_0^u f_1(v, I) \, dv \tag{73}$$

$$G_1(u, I) = 1 - \int_0^u F_1(v, I) \, dv \tag{74}$$

by numerical integration and plotting F_1 versus G_1 as shown by the solid line in Fig. 6. The curve starts at the point (1, 1), corresponding to $u = 0$, and curves in towards the origin as u increases, eventually crossing the G axis as shown. With the (G_1, F_1) curve as a guide, we draw in the (G, F) curve shown by the dashes. This curve is drawn so as to coincide with the solid curve in the region around (1, 1) and then depart from it in a "reasonable" manner so as to pass through the origin. Once the (G, F) curve is determined, the relation between G and u may be
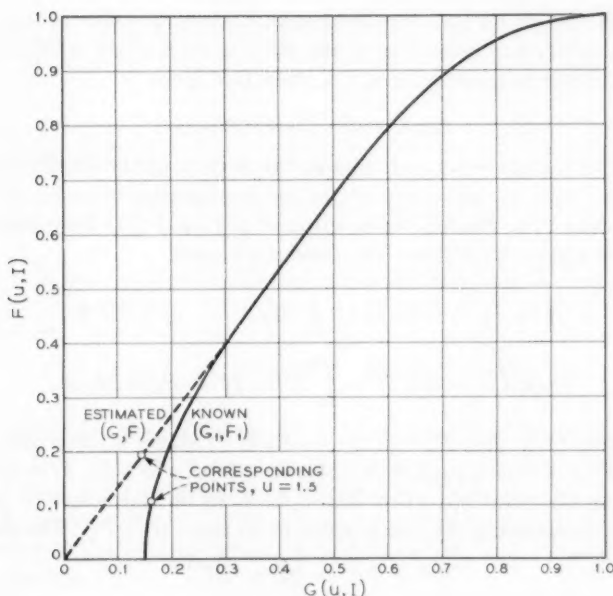
Fig. 6 — Method used to estimate probability distribution $F(u, I)$ for large values of u. These curves are for $I = 0$ and the normal-law power spectrum (16).

determined by numerical integration from

$$u = \int_g^1 \frac{dG}{F(G)} \tag{75}$$

and from this the cumulative probability $F(u, I)$ may be obtained as a function of u.

The justification of this procedure follows from the fact that functions $F(u, I)$, $G(u, I)$, computed from $f(u, I)$ (assumed known for the moment) by omitting the subscripts in (73) and (74), define a curve on Fig. 6 which starts at $(1, 1)$ for $u = 0$ and reaches the origin at $u = \infty$. Thus

$$F(\infty, I) = 1 - \int_0^\infty f(v, I)\, dv = 0,$$

$$G(\infty, I) = 1 - \int_0^\infty F(v, I)\, dv \tag{76}$$

$$= 1 - \int_0^\infty v f(v, I)\, dv = 0,$$

since the area under $f(u, I)$ is unity and the average value of u is unity.

We mention two properties of the (G, F) curve. First, if $dF/dG = \alpha$ in a small region around the origin, then (75) gives

$$F = Ce^{-\alpha u}, \ u \text{ large}, \tag{77}$$

where C is independent of u. Hence, the slope α of the (G, F) curve at the origin tells us something about the exponential behavior of F for large values of u. Second, subtraction of (73) and (74) from the corresponding equations without the subscripts gives

$$F(u, I) - F_1(u, I) = \int_0^u [f_1(v, I) - f(v, I)] \, dv$$

$$G_1(u, I) - G(u, I) = \int_0^u [F(u, I) - F_1(u, I)] \, dv. \tag{78}$$

Since $f_1(u, I) \geqq f(u, I)$, $F(u, I) - F_1(u, I)$ starts at zero and never decreases as u runs from 0 to ∞. The same is true of $G_1(u, I) - G(u, I)$. Therefore the point (G, F) on Fig. 6 tends to lie above and to the left of the corresponding (i.e. same value of u) point (G_1, F_1). The distance

$$[(F - F_1)^2 + (G - G_1)^2]^{1/2}$$

between the points is zero at $u = 0$ and never decreases as u runs to ∞.

One of the things to be decided in drawing the (G, F) curve is where it should start to diverge appreciably from the (G_1, F_1) curve. In drawing our curves we took this to be the region where $f_1(u, I)$ began to approach its constant value of $[1 - P(I)]/2$. Incidentally, it is helpful to draw the curves on both linear and log-log graph paper. The log-log paper is especially useful as a reminder that, after all, we are only making an estimate of the behavior of the distribution for large values of u.

Although $G(u, I)$ has been introduced here primarily for the purpose of estimating $F(u, I)$ for large values of u, it does have a significance of its own. Thus, let an instant be picked at random from the portion of time $I(t)$ exceeds I. This instant will fall in one of the $I(t) > I$ intervals. Let l be the distance between the instant and the end (or beginning) of the interval. Then $G(\tau/\bar{l}, I)$ is the chance that l exceeds τ. This is equivalent to a result given by C. Palm,[10] and may be obtained as follows:*

---

* Probabilities of the sort illustrated by expressions (79) and (81) have been considered by Siegert in connection with Markoff processes. These are discussed in Section IX. A closely related result due to A. Kohlenberg, M.I.T. Lincoln Lab. Memo. 44, Oct. 1953, has been generalized by McFadden.[7]

Suppose we observe $I(t)$ over a long period T. Let an instant be picked at random from the portion of time $I(t) > I$. The chance that l lies between τ and $\tau + d\tau$ is equal to

(number of $I(t) > I$ intervals longer than τ)$(d\tau)$
$$\frac{}{\text{total time } I(t) > I \text{ in period } T} \tag{79}$$
$$= \frac{(N_I T) P(\tau, I) \, d\tau}{(N_I T)\bar{l}} = P(\tau, I) \, d\tau/\bar{l} = F(u, I) \, du$$

where N_I is the average number of times per second $I(t)$ cuts upward across I and

$$P(\tau, I) = \int_\tau^\infty p(t, I) \, dt = \int_u^\infty f(v, I) \, dv = F(u, I) \tag{80}$$

is the probability that when an $I(t) > I$ interval is selected at random from the universe of such intervals its length will exceed τ. The chance that l will exceed τ is obtained by integrating (79) from τ to ∞, and is $G(\tau/\bar{l}, I)$, as we wished to show. The same sort of reasoning shows that the chance of the random instant falling in an $I(t) > I$ interval whose length exceeds τ is

$$\int_\tau^\infty t \left[-\frac{d}{dt} N_I T P(t, I) \right] dt/(N_I T\bar{l}) = (\tau/\bar{l}) F(\tau/\bar{l}, I) + G(\tau/\bar{l}, I) \tag{81}$$

where the right-hand side is obtained by integrating the left-hand side by parts.

Two further items concerning the problem of determining $p(\tau, I)$ for large values of τ are of interest. The first is that a second approximation to $p(\tau, I)$ is given by

$$p_2(\tau, I) = -2\pi\beta^{-1/2} e^{I^2/2} \int_0^\infty dI_1' \int_{-\infty}^0 dI_2' \int_I^\infty dI_3 \cdot$$
$$I_1' I_2' p(I, I_1', I_2', I, I_3) \tag{82}$$

which is much like (38) with the addition of the condition that $I_3 = I(t_3) > I$ where t_3 lies midway between t_1 and t_2. The integral is difficult to evaluate except when $I = 0$. Then we have*

---

* This result has been derived independently by S. Ehrenfeld and R. Goodman.[11] They have also studied the problem of computing the next case, in which $I(t) > 0$ at two points between t_1 and t_2.

$$p_2(\tau, 0) = \frac{N^2}{4\pi(\beta N_{55})^{1/2}N_{22}N_{33}D_3}$$
$$\left[\frac{1 + a - b - c}{1 + a} - \frac{a - bc}{D_3^{1/2}}(\alpha_1 + \alpha_2 + \alpha_3 - \pi)\right] \qquad (83)$$

where N_{ij} is the cofactor of the element ij in the determinant

$$N = \begin{vmatrix} 1 & 0 & m' & m & n \\ 0 & \beta & -m'' & -m' & -n' \\ m' & -m'' & \beta & 0 & n' \\ m & -m' & 0 & 1 & n \\ n & -n' & n' & n & 1 \end{vmatrix}$$

with n and n' being the values of m and m' computed for $\tau/2$. The other parameters are

$$a = -N_{23}(N_{22}N_{33})^{-1/2}, \qquad b = -N_{35}(N_{33}N_{55})^{-1/2},$$
$$c = N_{25}(N_{22}N_{55})^{-1/2},$$

$$\alpha_1 = \cot^{-1}\frac{a - bc}{D_3^{1/2}}, \qquad D_3 = \begin{vmatrix} 1 & c & b \\ c & 1 & a \\ b & a & 1 \end{vmatrix},$$

where $0 < \alpha_1 < \pi$ and α_2 and α_3 are obtained from α_1 by a cyclic permutation of a, b, c. It turns out that $b = c$ and $\alpha_2 = \alpha_3$. From the determinant N it may be shown that

$$N_{22} = N_{33} = \beta A + B - m'^2(1 - n^2),$$
$$N_{23} = m''A + B - m'^2(n^2 - m),$$
$$N_{25} = -N_{35} = [n'(1 + m) - nm']C,$$
$$N_{55} = [(\beta + m'')(1 + m) - m'^2]C,$$
$$N = (N_{22}^2 - N_{23}^2)/A,$$

where

$$A = (1 - m)(1 + m - 2n^2), \qquad B = n'(1 - m)[2m'n - n'(1 + m)],$$
$$C = (\beta - m'')(1 - m) - m'^2.$$

The second item is concerned with the approximate integral equation

$$p_1(\tau, I) = p(\tau, I) + \int_0^\tau p(y, I) q_1(\tau - y, I) \, dy \tag{84}$$

where $q_1(\tau, I) \, d\tau$ is the chance that $I(t)$ will pass downward through level I in $(\tau, \tau + d\tau)$, given that $I(t)$ passes downward through level I at $\tau = 0$. This is an extension of an equation given by D. S. Palmer.[5] When $I = 0$, (84) becomes McFadden's integral equation mentioned in the discussion following equation (19). Equation (84) is based on the assumption that the length of an $I(t) > I$ interval and the distance between the following downward crossings are independent. Changing the upper limit of integration for I_1' from ∞ to $-\infty$ on the right-hand side of (38) gives an expression for $q_1(\tau, I)$ which leads to

$$q_1(\tau, I) = p_1(\tau, I) + \left\{ M_{22} \beta^{-1/2} (1 - m^2)^{-3/2} \exp\left[\frac{I^2}{2} - \frac{I^2}{1 + m} \right] \right\}$$
$$\left\{ \frac{h}{\sqrt{2\pi}} \exp(-h^2/2) - \frac{(r + h^2)}{2} [1 - P(h)] \right\} \tag{85}$$

in which the various parameters have the same meaning as in (47). When $I = 0$, the results (58) lead to

$$q_1(\tau, 0) = M_{22} \beta^{-1/2} (1 - m^2)^{-3/2} [s - r \cos^{-1} r] / 2\pi.$$

VII. DISTRIBUTION OF INTERVAL LENGTHS FOR ENVELOPE

Here we study the envelope $R(t)$ of $I(t)$ in much the same way as $I(t)$ was studied in Section V. $I(t)$ is now taken to be a narrow-band Gaussian noise current whose power spectrum $w(f)$ is assumed to be symmetrical about the midband frequency f_0. We shall take $p(\tau, R)$ to be the probability density of the interval lengths during which $R(t)$ remains less than R. Strictly speaking, we should put some sort of a subscript, such as the r used on F in Section III, on $p(\tau, R)$ to distinguish it from the $p(\tau, I)$ of Section V. However, we shall not do so because the chance of confusion is small.

The first approximation to $p(\tau, R)$ is $p_1(\tau, R)$, where $p_1(\tau, R) \, d\tau$ is the chance that $R(t)$ will pass upwards through level R in $(t_2, t_2 + d\tau)$ given that $R(t)$ has passed downward through R at t_1, τ being the difference $t_2 - t_1$. Reasoning similar to that leading to (38) gives

$$p_1(\tau, R) = -\left(\frac{2\pi}{\beta} \right)^{1/2} R^{-1} e^{R^2/2} \int_{-\infty}^0 dR_1'$$
$$\int_0^\infty dR_2' \, R_1' \, R_2' \, p(R, R_1', R_2', R) \tag{86}$$

where now $b = 1$ and β is the β_r of (8). The function $p(R_1, R_1', R_2', R_2)$ is the probability density of $R(t_1)$, $R'(t_1)$, $R'(t_2)$, $R(t_2)$ and the primes denote time derivatives.

Let $I_c(t)$ and $I_s(t)$ denote the "in-phase" and "quadrature" components associated with $R(t)$. Let I_{c1}, I_{c1}', I_{c2}', I_{c2} stand for the values of $I_c(t)$ and $I_c'(t)$ at t_1 and t_2, and similarly for $I_s(t)$. Because of the symmetry of $w(f)$ about f_0, the I_c's are independent of the I_s's and their joint distribution is

$$p(I_{c1}, I_{c1}', I_{c2}', I_{c2})p(I_{s1}, I_{s1}', I_{s2}', I_{s2}), \tag{87}$$

where these two functions may be obtained from (39) by adding the subscripts c and s, respectively, to the I_1's and I_2's. The β and m appearing in the expression (40) for M are now given by

$$m = \int_0^\infty w\,(f)\cos 2\pi(f - f_0)\tau\, df, \tag{88}$$

$$\beta = [-m'']_{\tau=0}$$

instead of (41). These statements follow from results given in Appendix II of Ref. 3.

When we introduce the polar coordinates

$$I_{c1} = R_1 \cos \theta_1 \qquad I_{c2} = R_2 \cos \theta_2$$
$$I_{s1} = R_1 \sin \theta_1 \qquad I_{s2} = R_2 \sin \theta_2 \tag{89}$$

and their associated time derivatives, and then set

$$\varphi = \theta_1 - \theta_2, \qquad c = \cos \varphi, \qquad s = \sin \varphi,$$
$$R_1 = R_2 = R, \tag{90}$$

the joint distribution (87) goes into

$$\begin{aligned}
\frac{(2\pi)^{-4}}{M} \exp\Big\{ &-\frac{1}{2M} [(2M_{11} + cM_{14})R^2 + 2(M_{12} + cM_{24})(RR_1' - RR_2') \\
&+ M_{22}(R_1'^2 + R_2'^2) + 2cM_{23}R_1'R_2' \\
&+ M_{22}R^2(\theta_1'^2 + \theta_2'^2) + 2cM_{23}R^2\theta_1'\theta_2' \\
&+ 2sM_{23}R(R_1'\theta_2' - R_2'\theta_1') - 2sM_{24}R^2(\theta_2' + \theta_1')] \Big\}.
\end{aligned} \tag{91}$$

The quantity $p(R, R_1', R_2', R)dR_1 dR_2 dR_1' dR_2'$ is obtained by multiplying (91) by

$$dI_{c1} \cdots dI_{s2} = R^4\, dR_1\, dR_2\, dR_1'\, dR_2'\, d\theta_1\, d\theta_2\, d\theta_1'\, d\theta_2'$$

and integrating θ_1, θ_2 from 0 to 2π, and θ_1', θ_2' from $-\infty$ to $+\infty$. One

of the integrations with respect to the θ's may be performed. We write the other as an integration with respect to φ. When the integrations with respect to θ_1' and θ_2' are performed $p(R, R_1', R_2', R)$ is found to be

$$\left(\frac{R}{2\pi}\right)^2 \int_0^{2\pi} (M_{22}^2 - c^2 M_{23}^2)^{-1/2} e^A \, d\varphi \tag{92}$$

where A contains terms of the first and second degree in R_1', R_2'. Incidentally, the linear terms enter only in the combination $R_1' - R_2'$. Upon setting R_1', R_2' equal to $x_1 - a_1$, $x_2 - a_2$, respectively, and choosing a_1, a_2 so as to make the coefficients of x_1, x_2 in the resulting expression for A vanish we find

$$a_2 = -a_1, \tag{93}$$

$$(M_{22}^2 - M_{23}^2)a_1 = R\{M_{12}M_{22} + M_{23}M_{24} + c[M_{23}M_{12} + M_{22}M_{24}]\}.$$

The expressions given in (123) show that both sides of the equation for a_1 contain the factor M, which may be divided out with the result

$$a_1 = \frac{Rm'(c - m)}{1 - m^2}. \tag{94}$$

This choice of a_1 and a_2 and the results of Appendix I enable us to reduce (after a considerable amount of algebra) the expression for the exponent A in (92) to

$$-\frac{(1 - m^2)}{2(M_{22}^2 - c^2 M_{23}^2)} [M_{22}(x_1^2 + x_2^2) + 2cM_{22}x_1 x_2] - \frac{R^2(1 - mc)}{1 - m^2}. \tag{95}$$

When (92) is put in the expression (86) for $p_1(\tau, R)$ the result is a triple integral. The expression (95) suggests that we replace the variables R_1', R_2' by x, y defined by [see (46)]

$$x = -x_1[(1 - m^2)/M_{22}]^{1/2}, \qquad y = x_2[(1 - m^2)/M_{22}]^{1/2}, \tag{96}$$

$$\rho = cM_{23}/M_{22}.$$

Then (86) goes into

$$p_1(\tau, R) = \frac{RM_{22}e^{R^2/2}}{(2\pi\beta)^{1/2}(1 - m^2)^2} \int_0^{2\pi} J(\rho, k) \exp\left[-\frac{R^2(1 - mc)}{1 - m^2}\right] d\varphi \tag{97}$$

where $J(\rho, k)$ is obtained from (48) by putting ρ, k for r, h and

$$k = -a_1[(1 - m^2)/M_{22}]^{1/2}$$

$$= \frac{Rm'(m - c)}{1 - m^2}\left(\frac{1 - m^2}{M_{22}}\right)^{1/2}, \tag{98}$$

in which c stands for $\cos \varphi$. Here M_{22} and M_{23} are obtained by setting the values (88) for m and β in the appropriate expressions given in (120).

When τ becomes very large, k and ρ approach zero and we see that $p_1(\infty, R) = R(\beta/2\pi)^{1/2} \exp(-R^2/2)$ as it should according to (9). When τ approaches zero, with R fixed, the results of Appendix II enable us to show that

$$p_1(\tau, R) \rightarrow \frac{\tau R B}{2\beta\sqrt{2\pi}} \int_0^\infty J[1, R\beta B^{-1/2}(1 - x^2)]e^{-R^2x^2/2} \, dx$$

where $J(1, h)$ is given by (60).

The integral (97) must be evaluated numerically. However, a special case of interest in fading is obtained when R approaches zero. In this case, $p_1(\tau, R)$ tends to depend on τ and R only through the ratio τ/R. When we let τ and R both approach zero in such a way as to keep τ/R fixed, the results of Appendix II show that

$$\frac{M_{22}}{(1 - m^2)^2} \rightarrow \frac{B}{4\beta}, \qquad \frac{R^2(1 - mc)}{1 - m^2} \rightarrow \frac{R^2(1 - c)}{\beta\tau^2}, \tag{99}$$

$$k \rightarrow -R(1 - c)2B^{-1/2}\tau^{-2}.$$

Therefore, k approaches $-\infty$ except when φ is zero, and $J(\rho, k)$ is approximately k^2 except in a negligibly small region around $\varphi = 0$. It is convenient to introduce the variable $u = \tau/\bar{l}$, which, in the present case, becomes, with the help of the approximation (11) for $\bar{l}$,

$$u \rightarrow \frac{2\tau}{R} \left(\frac{\beta}{2\pi}\right)^{1/2}. \tag{100}$$

The function of u corresponding to $p_1(\tau, R)$ is

$$f_1(u, R) = \bar{l}p_1(\tau, R), \tag{101}$$

and it follows from (99), (100) and (101) that as R approaches zero,

$$f_1(u, R) \rightarrow \frac{z^2}{2} \int_0^{2\pi} (1 - c)^2 \exp[-z(1 - c)] \, d\varphi, \tag{102}$$

$$z = 2/(\pi u^2).$$

When we write $(1 - c)^2$ as $s^2 - 2c(1 - c)$ and integrate the s^2 portion by parts, we find that (102) may be expressed as a derivative:

$$f_1(u, R) \rightarrow -\frac{d}{du} \frac{1}{\pi u} \int_0^{2\pi} c \exp[-2\pi^{-1}u^{-2}(1 - c)] \, d\varphi$$

$$= -\frac{d}{du} \left[\frac{2}{u} I_1(z)e^{-z}\right] \tag{103}$$

where $I_1(z)$ is a Bessel function of imaginary argument. It may be shown that the expression under the differentiation sign is unity when $u = 0$ and is zero when $u = \infty$. Hence, an argument similar to the one following equation (65) suggests that as R approaches zero the approximation $p_1(\tau, R)$ approaches the actual probability density $p(\tau, R)$. Then $f_1(u, R)$ approaches $f(u, R)$ and (103) is the limiting form of both. We also have

$$F(u, R) \rightarrow \frac{2}{u} I_1(2/\pi u^2) \exp(-2/\pi u^2) \qquad (104)$$

where $F(u, R)$ is the probability that an interval during which $R(t) < R$ has a (normalized) length greater than u. $F(u, R)$ is the function denoted by $F_r(u, R)$ in Section III where equation (104) appears as equation (23). When u is very small or very large we have

$$u \rightarrow 0, \qquad\qquad u \rightarrow \infty,$$
$$f(u, R) \rightarrow 3\pi u/8, \qquad f(u, R) \rightarrow 6/\pi u^4, \qquad (105)$$
$$F(u, R) \rightarrow 1 - \tfrac{3}{16}\pi u^2, \qquad F(u, R) \rightarrow 2/\pi u^3.$$

The indefinite integral of the right-hand side of (104) is a constant plus

$$e^{-z}[I_0(z) + I_1(z)], \qquad z = 2/\pi u^2. \qquad (106)$$

Equation (97) refers to the intervals for which $R(t) < R$. If we are interested in the intervals for which $R(t) > R$, the analogue $p_1^*(\tau, R)$ of $p_1(\tau, R)$ is obtained from (86) by changing the signs of the ∞'s in the limits of integration. Instead of $J(\rho, k)$ in the integrand of (97) we get a double integral similar to (48) except that the upper limits of integration are $-\infty$'s instead of ∞'s. This changes the $J(\rho, k)$ to $J(\rho, -k)$. Upon using (56), we obtain

$$p_1^*(\tau, R) = p_1(\tau, R)$$
$$+ \frac{RM_{22}e^{R^2/2}}{(2\pi\beta)^{1/2}(1 - m^2)^2} \int_0^{2\pi} \left[(\rho + k^2)P(k) + \frac{2k}{\sqrt{2\pi}} e^{-k^2/2} \right] \cdot \qquad (107)$$
$$\exp\left[-\frac{R^2(1 - mc)}{1 - m^2} \right] d\varphi$$

which is the analogue of (57). When R becomes very large and τ small in such a way to keep $R\tau$ fixed, $p_1(\tau, R)$ approaches zero from physical considerations. We also have near $\varphi = 0$

$$\frac{R^2}{2} - \frac{R^2(1 - mc)}{1 - m^2} \rightarrow -\frac{\beta R^2\tau^2}{8} - \frac{R^2\varphi^2}{2\beta\tau^2} \qquad (108)$$
$$k \rightarrow R\beta B^{-1/2} - R\varphi^2\tau^{-2}B^{-1/2}.$$

The first line in (108) shows that most of the contribution to the integral in (107) comes from the region where φ is $O(\tau/R)$. In this region, the second term in the approximation for k is negligible in comparison with the first, so k is large and positive and hence $P(k)$ is nearly unity. Thus when R is large (107) becomes

$$p_1{}^*(\tau,\ R) \rightarrow \frac{\beta\tau R^2}{4}\ e^{-R^2\tau^2\beta/8} \tag{109}$$

which is of the same form as (66). Therefore, as we might expect, the interval lengths for large values of R are distributed in much the same way as are the interval lengths of $I(t)$ for large values of I.

VIII. COMPUTATIONS

The computations for the curves shown in Figs. 1 and 2 were based on either the power spectrum (16) or (20), depending on whether $I(t)$ or $R(t)$ was being considered. In both cases the bandwidth parameter σ was taken to be $1/(2\pi)$, so that both β and β_r were unity and

$$m = \exp\ (-\tau^2/2). \tag{110}$$

This autocorrelation function was also used by Palmer[5] in his computations.

Fig. 7 shows curves of the first approximation $f_1(u, I)$ to the probability density $f(u, I)$ for the lengths of the intervals during which $I(t)$
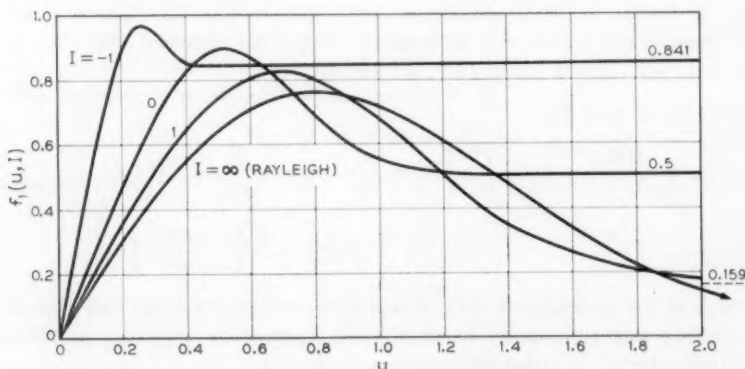

Fig. 7 — Graphs of the first approximation $f_1(u,\ I)$ when the power spectrum has the normal-law form (16). The rms value of $I(t)$ is one.
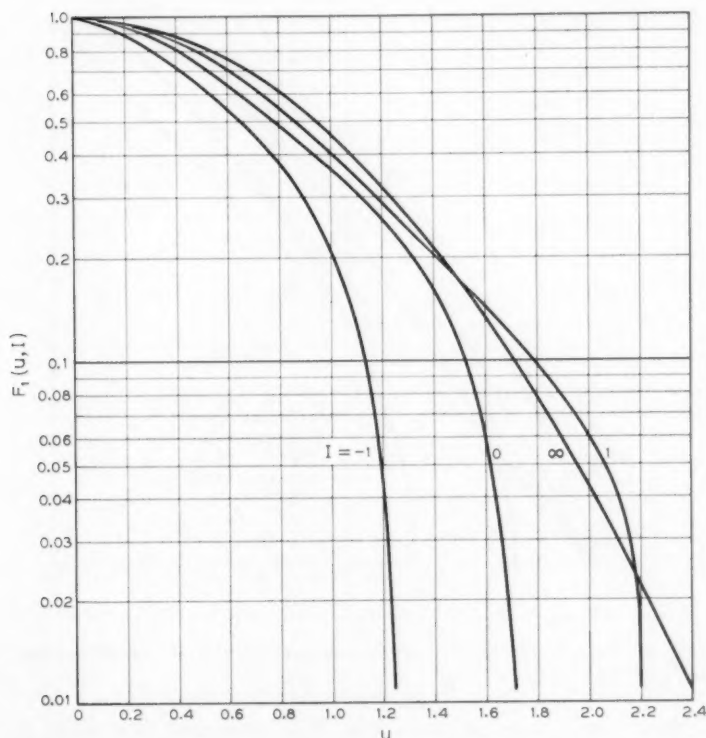
Fig. 8 — Graphs of the first approximation $F_1(u, I)$ to the interval length distribution $F(u, I)$. Estimates of $F(u, I)$ are shown in Fig. 1.

exceeds I. These curves were calculated from (47), (71) and tables[8] giving values of $K(r, h)$. Fig. 8 gives curves of $F_1(u, I)$ computed from $f_1(u, I)$ by means of (73). $F_1(u, I)$ is the first approximation to the cumulative distribution $F(u, I)$, estimates of which are shown in Fig. 1. One way of estimating $F(u, I)$ from $F_1(u, I)$ is to connect the $F_1(u, I)$ curve smoothly to a curve which approaches a straight line on the semi-log coordinates of Fig. 8, the slope and position of the straight line being chosen so as to make

$$\int_0^\infty F(u, I) \, du = 1. \qquad (111)$$

This relation follows from the fact that the average value of u is unity. The linear behavior ensures that $F(u, I)$ decreases exponentially as u
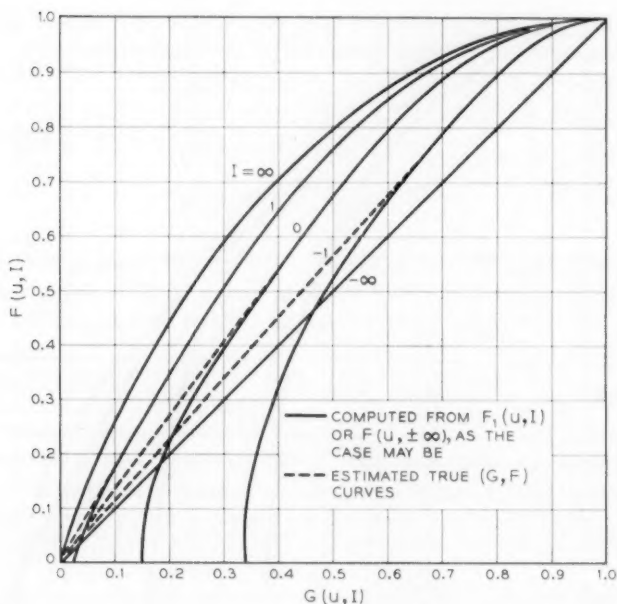
Fig. 9 — Graphs of (G_1, F_1) curves and estimates of the corresponding true (G, F) curves for $I(t)$.

becomes large. However, here $F(u, I)$ is estimated from $F_1(u, I)$ by the method outlined in Section VI. Fig. 9 shows (G, F) curves of the type illustrated in Fig. 6. The estimates of $F(u, I)$ given by Fig. 1 were obtained by connecting the curves of Fig. 9 to lines passing through the origin in a "smooth and reasonable" manner. The relation between u and $F(u, I)$ was determined by numerical integration from (75).

As a guide to what "smooth and reasonable" means, the case $I = 0$ was studied more completely with the help of the second approximation (83) and the computations made by Palmer[5] from his approximate integral equation. Some of this information is shown in Fig. 10. Palmer's results lead to a (G, F) curve (not shown) which lies slightly above the second approximation curve when G lies between 0.25 and 0.55. As G decreases below 0.25 the separation between the two curves increases and the curve from Palmer's results coincides with our estimate of the true (G, F) curve.

Fig. 11 shows curves of the first approximation $f_1(u, R)$ to the probability density $f(u, R)$ for the lengths of the intervals during which the
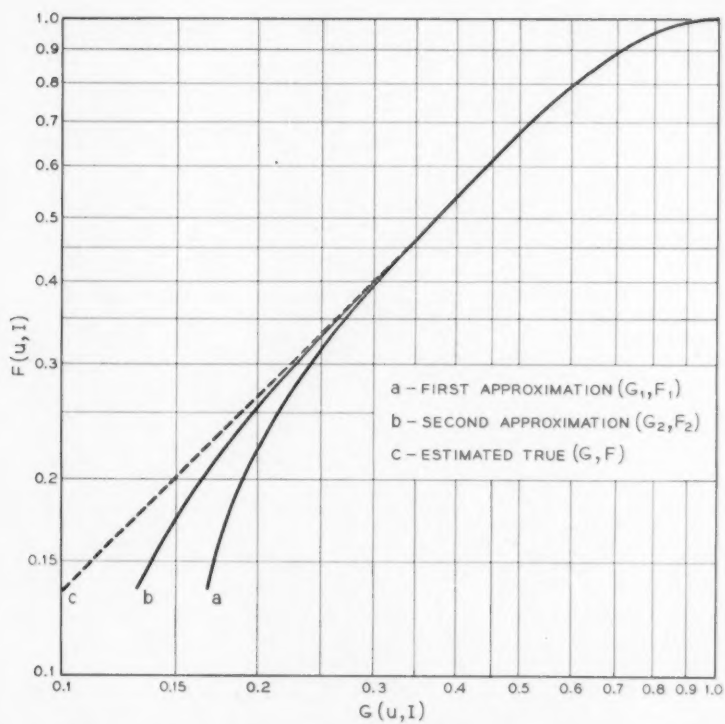
Fig. 10 — First and second approximation (G, F) curves for $I = 0$.
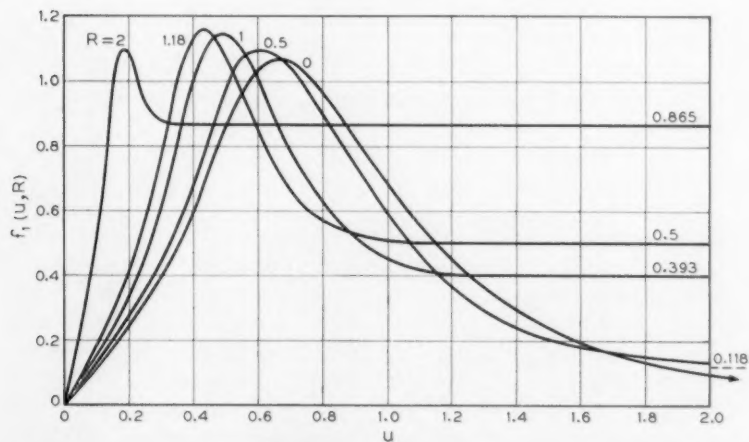

Fig. 11 — Graphs of the first approximation $f_1(u, R)$ when the power spectrum has the normal-law form (20). The median value of $R(t)$ is 1.18.
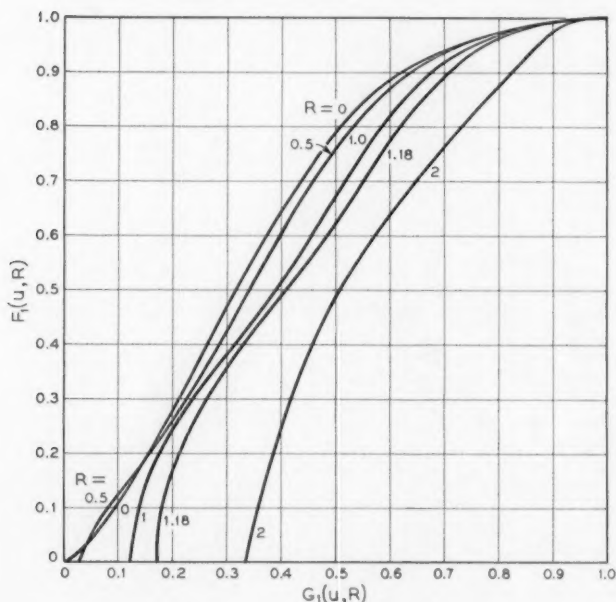
619

Fig. 12 — Graphs of the (G_1, F_1) curves for $R(t)$.

envelope $R(t)$ is less than R. These curves are the analogues of the curves for $I(t)$ shown in Fig. 7. Likewise, the (G, F) curves shown in Fig. 12 are the analogues of the curves of Fig. 9, and the estimates of $F_r(u, R)$ given in Table III were obtained from them. The functions $f_1(u, R)$, defined by (101), were computed by numerical integration from (97).

IX. COMMENTS ON A MARKOFF PROCESS

The foregoing work supposes that the autocorrelation function m of $I(t)$ is such that its second derivative with respect to τ exists at $\tau = 0$. This is not true when $w(f)$ has the form $4/[1 + (2\pi f)^2]$. In this case m is exp $(-|\tau|)$ and m'' is infinite at $\tau = 0$. $I(t)$ is now a "Markoff process" and a formal calculation of the average distance between successive zeros gives $\bar{l} = 0$.

Although the distribution function for the intervals between zeros does not exist in the usual sense, some related information is available. Most of it is due to A. J. F. Siegert.[12] Here we shall list several of his

results and comment on their relation to the work of this paper.* For the sake of simplicity we take $\overline{I^2(t)} = 1$ and $m = \exp(-|\tau|)$.

1. The probability $W(I, \tau)\,d\tau$ that $I(t)$ reaches the value 0 for the first time between τ and $\tau + d\tau$ ("first passage time") starting from I at time $t = 0$ is

$$W(I, \tau)\,d\tau = (2/\pi)^{1/2}I\,\exp(-I^2 z^2/2)(-dz),$$

$$z = e^{-\tau}(1 - e^{-2\tau})^{-1/2}, \qquad -dz = e^{2\tau}z^3\,d\tau. \qquad (112)$$

Wang and Uhlenbeck[13] obtained this result by solving an associated diffusion problem (by a method first used by Smoluchowski) in which the $I = 0$ axis corresponds to an absorbing barrier. Another derivation of (112) has been given by Siegert,[14] who obtains it as a solution of an integral equation.

2. The chance that $I(t)$ will go through zero for the first time in τ, $\tau + d\tau$ given that $I(0) > 0$ is

$$Q(\tau)\,d\tau = (2/\pi)e^{-\tau}(1 - e^{-2\tau})^{-1/2}\,d\tau. \qquad (113)$$

By symmetry, the condition $I(0) > 0$ may be replaced by $I(0) \neq 0$. $Q(\tau)\,d\tau$ is also the probability that the interval between an instant selected at random and the next following zero of $I(t)$ has a length l lying between τ and $\tau + d\tau$.

3. The chance that l exceeds τ seconds is

$$\int_\tau^\infty Q(l)\,dl = \frac{2}{\pi}\sin^{-1} e^{-\tau} \qquad (114)$$

where the arc lies between 0 and $\pi/2$.

4. Let $I(t)$ be observed over a long period of time T. The expected number of spacings between successive zeros whose lengths exceed τ seconds is

$$N_T(\tau) = Q(\tau)T. \qquad (115)$$

Furthermore, the probability that an instant selected at random will fall in a space whose length exceeds τ is

$$\tau Q(\tau) + (2/\pi)\sin^{-1} e^{-\tau}. \qquad (116)$$

Items 2, 3, and 4 are due to Siegert[12] and were obtained by analysis similar to that given in his paper.[14] David Slepian has pointed out, in

---

* I am indebted to Mr. Siegert for a copy of his report, which contains the results listed here.

some unpublished work, that (114) may also be obtained from the probability

$$P_{++} = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} e^{-\tau} \qquad (117)$$

that both $I(0)$ and $I(\tau)$ be positive. His derivation runs as follows. The probability (114) is twice the probability that $I(t)$ be positive over an interval of length τ selected at random. This last probability is $P_{++} - P_{+0+}$ where P_{+0+} is the chance that $I(0) > 0$, $I(\tau) > 0$ and $I(t)$ is zero at one or more points in $0 < t < \tau$. Since the process is Markoffian and symmetrical about zero, consideration of the behavior of $I(t)$ between its first zero (if it has one) and τ shows that P_{+0+} is equal to P_{+0-}, i.e. to P_{+-}. The sum of P_{++} and P_{+-} is the probability that $I(0)$ be positive, which is $\frac{1}{2}$. We then have

$$2(P_{++} - P_{+0+}) = 2(P_{++} - P_{+-}) = 4P_{++} - 1 = \frac{2}{\pi} \sin^{-1} e^{-\tau}$$

and this gives (114). The expression (117) for P_{++} is readily obtained by integrating the joint probability density.

5. It should be possible to obtain (114) by letting n approach infinity in the expression for two times the probability that $I(0)$, $I(\tau/n)$, $\cdots$, $I(n\tau/n)$ all be positive. The expression which should approach (114) is

$$2\pi^{-(n+1)/2}(1 - \alpha^2)^{1/2} \int_0^\infty dx_1 \cdots \int_0^\infty dx_{n+1} \cdot$$
$$\exp\left[-x_1^2 - x_{n+1}^2 - (1 + \alpha^2)(x_2^2 + \cdots + x_n^2)\right. \qquad (118)$$
$$\left. + 2\alpha(x_1x_2 + x_2x_3 + \cdots + x_nx_{n+1})\right]$$

where α denotes $\exp(-\tau/n)$.

This method of obtaining (114) is appealing because of its directness. Unfortunately, it seems that no one has yet succeeded in carrying out the limiting process. Some idea of how the limit is approached may be obtained from Table IV.

The values shown in the columns for $n = 1$ and $n = 2$ were computed from the corresponding known expressions for (118), namely

$$1 - \frac{1}{\pi} \cos^{-1} e^{-\tau} = 2P_{++},$$

$$1 - (\tfrac{1}{2\pi})(2 \cos^{-1} e^{-\tau/2} + \cos^{-1} e^{-\tau}).$$

When τ is small these behave like $1 - 0.45 \sqrt{\tau}$ and $1 - 0.54 \sqrt{\tau}$

TABLE IV — VALUES OF EXPRESSIONS (118) AND (114)

τ	$\epsilon^{-\tau}$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	Value of (114)
		Value of (118)				
0	1	1	1	1	1	1
0.05	0.951	0.900	0.879	0.868	0.861	0.800
0.10	0.905	0.860	0.830	0.814	0.804	0.720
0.15	0.861	0.830	0.793	0.774	0.761	0.660
0.20	0.819	0.805	0.763	0.742	0.726	0.611
0.40	0.670	0.734	0.672	0.641	0.620	0.474
∞	0	0.500	0.250	0.125	0.0625	0

respectively, while (114) becomes $1 - 0.90 \sqrt{\tau}$. When $n = 4$ the five-fold integral may be reduced to

$$2\pi^{-1/2}(1 - \alpha^2)^{1/2} \int_0^\infty dx_3 [K(r, -h, -k)]^2 \exp\ [-(1 - \alpha^2)x_3^2] \quad (119)$$

where $K(r, h, k)$ is the d/N tabulated in Ref. 8, and

$$r = \alpha(1 + \alpha^2)^{-1/2}, \qquad h = 2^{1/2}\alpha x_3, \qquad k = rh.$$

When α is near unity it is helpful to set

$$K(r, -h, -k) = 1 - L,$$

$$L = 1 - K(r, h, k) - \tfrac{1}{2}P(h) - \tfrac{1}{2}P(k).$$

The contribution of the 1 in $1 - 2L + L^2$ to (119) is unity and the contribution of $-2L + L^2$ may be obtained by numerical integration. Replacing the square of $K(r, -h, -k)$ in (119) by

$$K(r, -h, -k)[1 + P(h)]/2$$

gives an expression for (118) when $n = 3$.

6. The probabilities stated in Items 2, 3, and 4 have already appeared in Section VI for the well-behaved cases in which $\bar{l}$ exists. Thus, from (79) the analogue of $Q(\tau)\,d\tau$ is $P(\tau, I)\,d\tau/\bar{l}$, or $F(u, I)\,du$, and the analogue of (114) is $G(u, I)$. The probability (81) that a point chosen at random falls in an interval longer than τ goes into the corresponding expression (116).

These analogies lead us to regard $Q(\tau)$ and its integral (114) as being irregular versions of F and G (with τ playing the role of u) for which a (G, F) curve may be plotted with the help of expressions (113) and (114). When this is done, it is seen that F approaches infinity as G approaches unity. However, the other parts of the curve behave in much the same manner as in the earlier cases.

X. ACKNOWLEDGMENT

I wish to express my thanks to a number of my associates for their helpful remarks and for references to pertinent material. I am especially indebted to A. L. Durkee and D. Slepian for discussions on the contents of Sections IV and IX, respectively, and to Miss Katherine Jillson for computing the various tables and curves.

APPENDIX I

Minors of the Determinant M

Here we list the values of the minors of the determinant M defined by equation (40) and state some of the relations between them.

$$M_{11} = M_{44} = \beta^2 - m''^2 - \beta m'^2,$$

$$-M_{12} = M_{34} = m'm'' + mm'\beta,$$

$$M_{13} = -M_{24} = m'^3 - m'\beta - mm'm'', \tag{120}$$

$$M_{22} = M_{33} = \beta(1 - m^2) - m'^2,$$

$$M_{14} = mm''^2 - m\beta^2 - m'^2m'',$$

$$M_{23} = m''(1 - m^2) + mm'^2,$$

$$M = [(\beta + m'')(1 + m) - m'^2][(\beta - m'')(1 - m) - m'^2], \tag{121}$$

$$\frac{M_{11} + M_{14}}{\beta + m''} = \frac{M_{12} + M_{24}}{m'} = \frac{M_{22} - M_{23}}{1 + m}$$

$$= (\beta - m'')(1 - m) - m'^2 \tag{122}$$

$$M_{22} + M_{23} = (1 - m)[(\beta + m'')(1 + m) - m'^2]$$

$$M = (M_{11} + M_{14})(1 + m) - m'(M_{12} + M_{24}).$$

We also have relations of the form

$$M_{22}^2 - M_{23}^2 = \begin{vmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{vmatrix} = (-)^{2+2+3+3} \begin{vmatrix} 1 & m \\ m & 1 \end{vmatrix} M$$

$$= (1 - m^2)M,$$

$$M_{12}M_{22} + M_{23}M_{24} = \begin{vmatrix} M_{12} & M_{13} \\ M_{32} & M_{33} \end{vmatrix} = (-)^{1+2+3+3} \begin{vmatrix} 0 & -m' \\ m & 1 \end{vmatrix} M$$

$$= -m'mM, \tag{123}$$

$$M_{23}M_{12} + M_{22}M_{24} = m'M,$$

$$M_{11}M_{33} - M_{13}^2 = (\beta - m'^2)M,$$

$$M_{14}M_{23} + M_{13}^2 = (m'^2 - mm'')M,$$

$$M_{11}M_{23} + M_{14}M_{22} = (m'' - \beta m)M.$$

APPENDIX II

Expressions for Various Parameters When τ Becomes Small

In dealing with $I(t)$ we shall put

$$b_{2n} = (2\pi)^{2n} \int_0^\infty w(f)f^{2n}\, df, \qquad n = 0, 1, 2, \cdots \qquad (124)$$

and in particular

$$b_0 = 1, \qquad b_2 = \beta, \qquad b_4 = \gamma,$$

$$B = b_4 - b_2^2 = \gamma - \beta^2, \qquad (125)$$

$$C = b_2 b_6 - b_4^2 = \beta b_6 - \gamma^2.$$

The results for $I(t)$ may be carried over directly into those for the envelope $R(t)$ of a narrow-band noise current whose power spectrum is symmetrical about the midband frequency f_0. This is accomplished by replacing the definition (124) for b_{2n} by

$$b_{2n} = (2\pi)^{2n} \int_0^\infty w(f)(f - f_0)^{2n}\, df. \qquad (126)$$

For the normal-law power spectrum used in our computations γ is equal to $3\beta^2$ and B becomes $2\beta^2$.

It follows from (41) that when τ is small and the b's exist,

$$m = 1 - \frac{\beta\tau^2}{2!} + \frac{\gamma\tau^4}{4!} - \frac{b_6\tau^6}{6!} + \cdots,$$

$$m' = -\beta\tau + \cdots, \qquad (127)$$

$$m'' = \, = -\beta + \cdots.$$

These expansions may be used to show that

$$1 - m \to \frac{\beta\tau^2}{2},$$

$$1 + m \to 2 - \frac{\beta\tau^2}{2},$$

$$1 - m^2 \to \beta\tau^2 - \cdots,$$

$$(\beta + m'')(1 + m) - m'^2 \to \tau^2 B,$$

$$(\beta - m'')(1 - m) - m'^2 \to \tau^6 C/144,$$

$$M_{23} \to M_{22} \to \tau^4 \beta B/4,$$

$$M_{22} - M_{23} \to \tau^6 C/72,$$

$$M \to \tau^8 BC/144,$$

$$M_{22}^2 - M_{23}^2 \to \tau^{10} \beta BC/144,$$

$$(1 - m^2)/M_{22} \to 4/B\tau^2.$$

$$(128)$$

APPENDIX III

An Approximation to the Cumulative Probability $P(\tau, I)$

In Section V we worked with a first approximation $p_1(\tau, I)$ to the probability density $p(\tau, I)$ for the lengths of the $I(t) > I$ intervals. Here we note a rather simple approximation $P_1(\tau, I)$ to the probability $P(\tau, I)$ that an interval selected at random from the universe of $I(t) > I$ intervals will be longer than τ. We take $P_1(\tau, I)$ to be the limit, as dt approaches zero, of a fraction having the numerator

[Probability that $I(\tau) > I$ given that $I(t)$ passes upward
through the value I during the interval $(0, dt)$]

and the denominator

[Probability $I(t)$ passes upward through I in $(0, dt)$].

By proceeding as in Section V it may be shown that

$$P_1(\tau, I) = \frac{1}{2} - \frac{1}{2} P\left(\frac{k}{s}\right) + \frac{r}{2}\left[1 + P\left(\frac{rk}{s}\right)\right] \exp\left(-k^2/2\right) \quad (129)$$

where $P(x)$ is the error integral (3), and now

$$r = -m'\beta^{-1/2}(1 - m^2)^{-1/2}, \qquad s^2 = 1 - r^2,$$

$$k = I[(1 - m)/(1 + m)]^{1/2},$$

with m and β the same as in Section V. The approximation given by $P_1(u\bar{l}, I)$ to $F(u, I)$ is not as good as the $F_1(u, I)$ given by (73) (which is based on the $p_1(\tau, I)$ of Section V), but we do have $P_1(u\bar{l}, I) \geqq F(u, I) \geqq F_1(u, I)$. This may lead to another proof of (66).

APPENDIX IV

Fade Length Distributions Having a Large Percentage of Long Fades

Some, but not all, of the small number of available experimental observations of fade lengths show a greater percentage of long fades than those shown in Figs. 2 and 3. Since our calculations are based on a normal-law power spectrum, it is of interest to know whether a different power spectrum may give a larger percentage of long fades. An example of such a power spectrum is studied in the first part of this appendix. The second part is concerned with a guess as to how the very long fades are distributed.

The power spectrum

$$w(f) = \frac{4}{1 + (2\pi f)^2}, \tag{130}$$

which has been discussed in Section IX, is quite different from the normal-law type. The power carried by the high frequencies in this spectrum is so large that $\bar{l} = 0$, and our method of investigating interval lengths breaks down. Our formulas will still apply, though, if the high frequencies are attenuated by some factor such as $a^4/[a^2 + (2\pi f)^2]^2$. An attenuation of this type might either exist in nature or be produced by the measuring apparatus. Here we shall take

$$w(f) = \frac{8a^3(a + 1)^2/(2a + 1)}{[1 + (2\pi f)^2][a^2 + (2\pi f)^2]^2} \tag{131}$$

where the numerator has been chosen so as to make the integral of $w(f)$ from 0 to ∞ equal to unity.

If it were not for the considerable amount of additional work required, we would investigate the interval length distribution associated with the envelope of an $I(t)$ having the power spectrum obtained from (131) by setting $(f - f_0)$ in place of f and changing the 8 in the numerator to
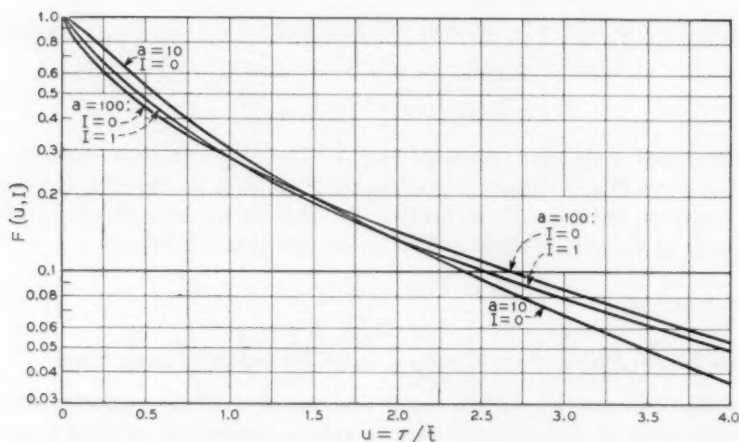
Fig. 13 — The probability $F(u, I)$ that $I(t) > I$ for an interval lasting longer than τ. $I(t)$ has the power spectrum given by equation (131) and $\bar{l}_0$ is $\pi a^{-1}\sqrt{2a+1}$.

4. Instead, we shall apply the results of Section V to obtain the interval length distribution associated with an $I(t)$ having the power spectrum (131), and rely on the similarity between the curves of Figs. 1 and 2.

The general theory of Section VII shows that, as R approaches 0, the fade length distribution $F_r(u, R)$ of the envelope approaches the form $F_r(u, R \to 0)$ shown in Fig. 4. However, when a is very large it appears that this limiting form will have to be approached by going through shapes similar to those shown in Fig. 13. This indicates that the values of R will have to be very small indeed before $F_r(u, R)$ begins to look like the $F_r(u, R \to 0)$ of Fig. 4.

The autocorrelation function m, defined by (41), and its derivatives with respect to τ are, assuming $\tau > 0$,

$$
\begin{aligned}
m &= x - y[3a^2 - 1 + a\tau(a^2 - 1)], \\
m' &= -x + ya[2a^2 + a\tau(a^2 - 1)], \\
m'' &= x - ya^2[a^2 + 1 + a\tau(a^2 - 1)],
\end{aligned}
\tag{132}
$$

where

$$
\begin{aligned}
x &= 2Aa^3 e^{-\tau}, \qquad y = A e^{-a\tau}, \\
A &= 1/(2a + 1)(a - 1)^2.
\end{aligned}
\tag{133}
$$

The expansion of m in powers of τ up to the fifth is

$$
m = 1 - \frac{\beta\tau^2}{2!} + \frac{\gamma\tau^4}{4!} - \frac{a^3 \mid \tau \mid^5}{5!} \frac{2(a + 1)^2}{2a + 1},
\tag{134}
$$

in which

$$\beta = a^2/(2a + 1), \qquad \gamma = a^3(a + 2)/(2a + 1),$$
$$B = \gamma - \beta^2 = 2a^3(a + 1)^2/(2a + 1)^2. \tag{135}$$

From β and γ it follows that the average interval between the zeros of $I(t)$ is $\pi a^{-1}\sqrt{2a + 1}$, and the expected number of maxima per second is $(2\pi)^{-1}\sqrt{a(a + 2)}$.

These expressions were used to compute $p_1(\tau, I)$ from (47), and the method outlined in Section VI was applied to obtain the curves shown in Fig. 13. These curves are for $a = 10$ and $a = 100$. Difficulty was encountered, in the $a = 100$ case, in extrapolating the distributions for small values of τ to obtain $F(u, I)$ for the larger values of u. Consequently, these curves are not too reliable, but there seems to be enough truth in them to show that the percentage of long fades is larger than for the distributions of Fig. 1 pertaining to the normal-law power spectrum.

Irrespective of whether power spectra of the form (131) have any practical significance, it is interesting to speculate on how the corresponding interval length distribution $F(u, I)$ behaves for large values of u. When a becomes large, $I(t)$ has many more maxima than zeros (the average number of maxima between successive zeros approaches $\sqrt{a/2}$), and we are led to picture the power spectrum (131) as consisting of two parts.

One part is the "low-frequency" portion which behaves like (130) and extends from zero up to a frequency at which the cutoff factor begins to operate. The other, the high-frequency portion, extends from this frequency to infinity. Most of the power is in the low-frequency part and we may regard it as producing most of the deviation of $I(t)$ from zero. The high-frequency part produces small rapid fluctuations which are superposed on the more slowly changing low-frequency portion of $I(t)$. Thus the high-frequency part produces the large number of maxima and the low-frequency part controls the drift towards or away from the value $I(t) = 0$. According to this picture some relation might be expected between $F(u, 0)$ and $Q(\tau)$ [given by (113)] when u is large. In fact, comparison of (79) and (113) suggests that $F(u, 0)\, du$ approaches $Q(\tau)\, d\tau$ for the larger values of τ:

$$F(u, 0) \to \bar{l}(2/\pi)e^{-\tau}(1 - e^{-2\tau})^{-1/2}, \qquad \tau = u\bar{l}. \tag{136}$$

The curves shown in Fig. 13 indicate that (136) may not hold when $a = 10$, but may do so when $a = 100$. Evidently, if (136) is to be valid, it is necessary that the number of maxima between zeros be large. This is

also suggested by the original derivation of (112), which is based upon an analogy with the Brownian motion of a particle subject to an elastic force.

If the relation (136) between $F(u, 0)$ and $Q(\tau)$ should turn out to be true, it would be helpful to have expressions for the generalizations $Q(\tau, I)$ and $Q_r(\tau, R)$ of $Q(\tau)$ corresponding to the "first passage times" down through the level $I(t) = I$ and up through the level $R(t) = R$, respectively. These generalizations may be obtained by the method used by Wang and Uhlenbeck mentioned in Item 1 of Section IX. The derivations are sketched in the following paragraphs.

Let $I(t)$ have the power spectrum (130) and let $W(I, I_1, \tau)d\tau$ be the probability that $I(t)$ passes down through the level I for the first time in the interval $\tau, \tau + d\tau$ when it starts from the level I_1, which is greater than I, at time $t = 0$. Then $W(I, I_1, \tau) \, d\tau$ is equal to $(\partial P/\partial x) \, d\tau$ evaluated at $x = I$ and $t = \tau$, where P is the solution of the diffusion equation

$$\frac{\partial P}{\partial t} = \frac{\partial^2 P}{\partial x^2} + x \frac{\partial P}{\partial x} + P \tag{137}$$

which is zero at $x = I$ for all time and is concentrated around $x = I_1$ at time 0 [i.e., equals $\delta(x - I_1)$ at time $t = 0$]. The probability $Q(\tau, I) \, d\tau$ that if $I(0) > I$ then $I(t)$ passes down through the level I for the first time in $\tau, \tau + d\tau$ is obtained from

$$Q(\tau, I) = \int_I^\infty \frac{\exp\,(-I_1{}^2/2)}{\sqrt{2\pi}} \, W(I, I_1, \tau) \, dI_1 / \operatorname{prob}[I(0) > I]. \tag{138}$$

A typical solution of (137) is

$$e^{-x^2/4} D_n(x) \, e^{-nt} \tag{139}$$

where $D_n(x)$ is the parabolic cylinder function. The boundary condition at $x = I$ gives the equation

$$D_n(I) = 0 \tag{140}$$

to determine the eigenvalues n_i, $i = 1, 2, \cdots$. A formal application of the Sturm-Liouville theory shows that P may be expressed as a sum of terms of the type (139), with suitable coefficients, when n is replaced by n_i and the summation taken over i. If the typical eigenfunction is denoted by $a_i D_{n_i}(x)$, the relation to determine the normalizing constant a_i is obtained by differentiating

$$(n - n_i) \int_I^\infty D_n(x) D_{n_i}(x) \, dx = -D_n(I) D_{n_i}{}'(I) \tag{141}$$

with respect to n and then setting $n = n_i$. In (141), which may be obtained from the differential equation for $D_n(x)$, we have denoted $\partial D_n(x)/\partial x$ by $D_n'(x)$.

It is found that

$$W(I, I_1, \tau) = \exp\,[(I_1^2 - I^2)/4] \sum_{i=1}^{\infty} a_i^2 D_{n_i}(I_1) D_{n_i}'(I)\, e^{-n_i \tau}, \quad (142)$$

$$Q(\tau, I) = \frac{(2\pi)^{-1/2}\, e^{-I^2/2}}{2^{-1}[1 - P(I)]} \sum_{i=1}^{\infty} \frac{D_{n_i}'(I)\, e^{-n_i \tau}}{n_i[-\partial D_n(I)/\partial n]_{n=n_i}}, \quad (143)$$

where $P(I)$ is the error integral (3) and

$$a_i^{-2} = D_{n_i}'(I)\, [-\partial D_n(I)/\partial n]_{n=n_i}. \quad (144)$$

The integration (138) used to obtain (143) from (142) makes use of a result obtained by integrating a form of the differential equation for $D_n(x)$.

Siegert's result (115) now reads as follows: When $I(t)$ is observed over a long period of time T, the expected number of $I(t) > I$ intervals whose lengths exceed τ seconds is

$$N_T(\tau, I) = \frac{T}{2}\, [1 - P(I)]Q(\tau, I). \quad (145)$$

This may be seen by writing (79) as

$$\frac{N_T(\tau, I)\, d\tau}{T \text{ prob } [I(t) > I]} = Q(\tau, I)\, d\tau.$$

When $I = 0$, the values of the n_i are 1, 3, 5, $\cdots$ and it may be shown that (143) reduces to (113) as it should. As I becomes large and positive the n_i's increase in value, so that $n_1 \approx I^2/4$ and the first few succeeding n_i's differ from n_1 by terms of order $I^{2/3}$. Furthermore, setting $n = n_i$ and approximating $D_{n-1}(I)$ by $(-1)\,\partial D_n(I)/\partial n$ in the relation

$$D_n'(I) + \frac{I}{2}\, D_n(I) - nD_{n-1}(I) = 0$$

suggests that when I is large and i not too large

$$D_{n_i}'(I)/n_i[-\partial D_n(I)/\partial n]_{n_i} \approx 1,$$

and hence we may expect $Q(\tau, I)$ to be of the form of $I \exp\,(-I^2\tau/4)$ times a more slowly varying function of τ and I.

In much the same way, expressions may be derived for $W_r(R, R_1, \tau)$ and $Q_r(\tau, R)$ corresponding to the envelope $R(t)$ of a Gaussian noise

current $I(t)$ whose power spectrum is obtained from (130) by substituting $f - f_0$ for f and replacing the 4 in the numerator by 2. Instead of (137) we must use

$$\frac{\partial P}{\partial t} = \frac{\partial^2 P}{\partial r^2} + \left(r - \frac{1}{r}\right)\frac{\partial P}{\partial r} + \left(1 + \frac{1}{r^2}\right)P \tag{146}$$

where $P = 2\pi r\, P_x P_y$ now is the probability density or concentration corresponding to a narrow circular ring of radius $r = \sqrt{x^2 + y^2}$ and P_x, P_y are, respectively, solutions of (137) and the corresponding equation in y.

The probability $W_r(R, R_1, \tau)\, d\tau$ that the envelope $R(t)$ passes up through the value R for the first time in τ, $\tau + d\tau$ when it starts from R_1, less than R, at time $t = 0$ is $-(\partial P/\partial r)\, d\tau$ evaluated at $r = R$ and $t = \tau$. Here P is the solution of (146) which is zero at the boundary $r = R$ for all t and is $\delta(r - R_1)$ at time $t = 0$.

The typical solution of (146) is

$$r^{1/2} e^{-r^2/4} v(\alpha, r)\, e^{-\alpha t} \tag{147}$$

where α is a parameter similar to n in the earlier case, and

$$v(\alpha, r) = r^{1/2}\, e^{-r^2/4}\, {}_1F_1\left(-\frac{\alpha}{2}; 1; \frac{r^2}{2}\right) \tag{148}$$

becomes the typical orthogonal function for the range $0 \leq r \leq R$ when α is one of the roots α_i, $i = 1, 2, 3, \cdots$ of the equation

$${}_1F_1\left(-\frac{\alpha}{2}; 1; \frac{R^2}{2}\right) = 0. \tag{149}$$

For very large values of R, α_i approaches $0, 2, 4, \cdots$ and the confluent hypergeometric function in (148) becomes the Laguerre polynomial $L_{i-1}(r^2/2)$. For very small values of R, $v(\alpha_i, r)$ approaches

$$\sqrt{r}\, J_0(r\sqrt{\alpha_i + 1})$$

and $R\sqrt{\alpha_i + 1}$ becomes the ith root of the Bessel function $J_0(x)$. Some rough work indicates that α_1 is approximately 5 for $R = 1$ and 3.25 for $R = 1.18$. When R is very small $\alpha_1 + 1$ is approximately $5.78/R^2$.

It is found that

$$P = (r/R_1)^{1/2} \exp\left[(R_1^2 - r^2)/4\right] \sum_{i=1}^{\infty} a_i^2 v(\alpha_i, R_1) v(\alpha_i, r)\, e^{-\alpha_i t} \tag{150}$$

where the normalizing constant a_i is given by

$$a_i^2 \left[\frac{\partial v(\alpha, R)}{\partial \alpha} \right]_{\alpha = \alpha_i} \left[\frac{\partial v(\alpha_i, r)}{\partial r} \right]_{r = R} = 1.$$

When the expression for $W_r(R, R_1, \tau)$ obtained by differentiating P is set in the equation corresponding to (138), we obtain

$$Q_r(\tau, R) = \frac{R}{\exp [R^2/2] - 1} \sum_{i=1}^{\infty} \frac{e^{-\alpha_i \tau}}{\alpha_i} \frac{[\partial v(\alpha_i, R)/\partial R]}{[\partial v(\alpha, R)/\partial \alpha]_{\alpha = \alpha_i}} \qquad (151)$$

where $Q_r(\tau, R) \, d\tau$ is the probability that $R(t)$ will pass up through the level R for the first time in $\tau, \tau + d\tau$, given that $R(0) < R$.

In the limit as R becomes small, (151) reduces to

$$Q_r(\tau, R) \to 4R^{-2} \sum_{i=1}^{\infty} e^{-\alpha_i \tau}, \qquad (152)$$

$$R\sqrt{\alpha_i} \to [i\text{th root of } J_0(x)].$$

When (152) is integrated with respect to τ between the limits 0 and ∞ the result is 1, as it should be. The result (152) has probably been given before, since it is closely related to Brownian motion of a particle and to the first passage time across a circle, but a suitable reference has not been found by the author.

When R is small so that prob $[R(t) < R] \to R^2/2$, the relation (152) and the analogue of (145) for $R(t)$ lead to

$$N_T(\tau, R) \to 2T \sum_{i=1}^{\infty} e^{-\alpha_i \tau} \qquad (153)$$

for the expected number, in the long time T, of $R(t) < R$ intervals whose lengths exceed τ seconds.

All of the discussion between equations (136) and (153) pertains to the power spectrum (130) or its analogue for $R(t)$. In the comparison of theory and experiment given in Section IV, it is convenient to deal instead with

$$w(f) = \frac{4k}{k^2 + (2\pi f)^2}. \qquad (154)$$

The analogue of (154) for $R(t)$ has already appeared as equation (24). Results for (154) pertaining to duration of intervals, time averages and so on may be obtained from the corresponding results for (130) by replacing t by kt. Thus, for example, the probability $Q_r(\tau, R) \, d\tau$ goes into

$kQ_r(k\tau, R)\, d\tau$ and the number of $R(t) < R$ intervals longer than τ to be expected in time T, when R is small, is

$$N_T(\tau, R) \to 2kT \sum_{i=1}^{\infty} e^{-\alpha_i k\tau}, \tag{155}$$

where α_i is still given by the second relation in (152). Siegert's expression (115) for the expected number of spacings between successive zeros whose lengths exceed τ seconds is now

$$N_T(\tau) = (2kT/\pi)\, e^{-k\tau}(1 - e^{-2k\tau})^{-1/2}.$$

Incidentally, this may possibly lead to an estimate of k when one is trying to measure a power spectrum which insists on increasing as $1/f^2$ at the lowest frequencies which can be conveniently measured. The problem of estimating k from $N_T(\tau)$ has been discussed by Siegert.[12]

If the "low-frequency" portion of a power spectrum behaves like (154) instead of (130) the remarks in the preceding paragraph show that the analogue of (136) is

$$F(u, 0) \to k\bar{l}(2/\pi)\, e^{-k\tau}(1 - e^{-2k\tau})^{-1/2} \tag{156}$$

where $u = \tau/\bar{l}$ is assumed to be very large. The analogous surmise for $F_r(u, R)$ is that

$$F_r(u, R) \to k\bar{l}Q_r(k\tau, R) \tag{157}$$

where the low-frequency behavior of the power spectrum is given by (24). When R is small, this and (152) suggest that

$$F_r(u, R) \to 4k\bar{l}R^{-2} \sum_{i=1}^{\infty} e^{-\alpha_i k\tau} \tag{158}$$

for the probability that a long fade below the level R will last longer than τ seconds. Multiplying both sides of (158) by N_R and using

$$N_R\bar{l} = \text{prob } [R(t) < R] \to R^2/2$$

shows that the expected number of fades below level R, in unit time, whose lengths exceed τ is

$$N_R F_r(u, R) \to 2k \sum_{i=1}^{\infty} e^{-\alpha_i k\tau} \tag{159}$$

where α_i is related to the zeros of $J_0(x)$ as indicated by (152). The relation (159) is to be expected in view of (155).

REFERENCES

1. Bullington, K., Inkster, W. J. and Durkee, A. L., Results of Propagation Tests at 505 mc and at 4090 mc on Beyond-Horizon Paths, Proc. I.R.E., **43,** October, 1955, pp. 1306–1316.
2. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., **24,** January, 1945, pp. 46–156.
3. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., **27,** January, 1948, pp. 109–157.
4. Norton, K. A., Rice, P. L., Janes, H. B. and Barsis, A. P., The Rate of Fading in Propagation Through a Turbulent Atmosphere, Proc. I.R.E., **43,** October, 1955, pp. 1341–1353; Norton, K. A., Vogler, L. E., Mansfield, W. V. and Short, P. J., The Probability Distribution of the Amplitude of a Constant Vector Plus a Rayleigh-Distributed Vector, Proc. I.R.E., **43,** October, 1955, pp. 1354–1361.
5. Palmer, D. S., Properties of Random Functions, Proc. Cambridge Phil. Soc., **52,** October, 1956, pp. 672–686.
6. Favreau, R. R., Low, H and Pfeffer, I., Evaluation of Complex Statistical Functions by an Analogue Computer, presented at the 1956 I.R.E. Convention. An abstract appears in Proc. I.R.E., **44,** March, 1956, p. 388.
7. McFadden, J. A., The Axis-Crossing Intervals of Random Functions, I.R.E. Trans. on Information Theory, **IT-2,** December, 1956, p. 146; also a paper scheduled for the March 1958 issue of Trans. on Information Theory.
8. Pearson, Karl, ed., *Tables for Statisticians and Biometricians*, Cambridge Univ. Press, 1931, Part II, Table VIII, Volumes of Normal Bivariate Surface, pp. 78–109.
9. Kuznetsov, Stratonovich and Tikhonov, On the Duration of Exceedances of a Random Function, Journal of Technical Physics, **24,** 1954. Translated from the Russian by Goodman, N. R., New York Univ. Coll. of Engineering, Scientific Paper No. 5, Engineering Statistics Group, March, 1956.
10. Palm, Conny, *Intensitätsschwankungen im Fernsprechverkehr*, Ericsson Technics 44, L. M. Ericsson, Stockholm, 1943.
11. Ehrenfeld, S. and Goodman, R., private communication.
12. Siegert, A. J. F., On the Roots of Markoffian Random Functions, Rand memorandum, September 5, 1950.
13. Wang, Ming Chen and Uhlenbeck, G. E., On the Theory of the Brownian Motion II, Revs. Modern Phys., **17,** April–July, 1945, p. 323.
14. Siegert, A. J. F., On the First Passage Time Probability Problem, Phys. Rev., **81,** February 15, 1951, pp. 617–623.

Frequency Shifts in Cavities with Longitudinally Magnetized Small Ferrite Discs

By H. SEIDEL and H. BOYET

(Manuscript received August 27, 1957)

Values of the tensor permeability components of a magnetized ferrite at microwave frequencies may be determined from the frequency shift that a sample produces in a resonant cavity.

In this paper mathematical expressions are obtained relating this frequency shift to the diagonal and off-diagonal permeability tensor values and to the cavity geometry and ferrite geometry for any TE, TM or TEM mode natural to the empty cavity. The cavities considered are axial and have generalized cross section. The expressions are valid for ferrite discs whose volume is small compared with the cavity volume; it is assumed that the ferrite sample does not perturb the fields outside the ferrite from their empty cavity values. The cross section of the ferrite disc is arbitrary in shape and is perpendicular to the axis of the cavity.

Examples are given in which the frequency shifts in a circular coaxial cavity, circular cylindrical cavity and rectangular cavity containing thin ferrite discs are derived for general TE and TM modes. The TEM-type cavity is also considered.

I. INTRODUCTION

A possible experimental determination of the elements of the permeability tensor of a magnetized ferrite at any frequency and magnetic field rests on the measurement of the frequency shift that the sample produces in a resonant cavity.[1] From this measurement and from the theoretical relation among frequency shift, ferrite and cavity geometry and ferrite properties, we may deduce the diagonal and off-diagonal components, μ and κ, of the Polder tensor.

The theoretical relation referred to above is the result of the assumption usually made that the ferrite has small volume compared to the

cavity volume and produces a small perturbation of the fields in the cavity:[1,2]

$$\frac{2\,d\omega}{\omega_0} = \frac{W_s^{(M)} + W_s^{(E)}}{W_c^{(E)}}.$$ (1)

Here W_c is the total energy stored in the empty cavity at resonance, $W_s^{(M)}$ the additional magnetic energy stored in the sample, $W_s^{(E)}$ the additional electric energy stored in the sample, $d\omega$ the shift of resonance frequency upon introduction of the sample and ω_0 the resonant frequency of the empty cavity. The quantities $W_c^{(E)}$, $W_s^{(M)}$ and $W_s^{(E)}$ are given by

$$W_c^{(E)} = \frac{\epsilon_0}{2} \int_{\text{cavity}} \mathbf{E}_0 \cdot \mathbf{E}_0^* d\tau;$$

$$W_s^{(M)} = \frac{\mu_0}{2} \int_{\text{sample}} \mathbf{M} \cdot \mathbf{H}_0^* d\tau; \qquad \mathbf{M} = \chi \mathbf{H}$$ (2)

$$W_s^{(E)} = \frac{\epsilon_0}{2} \int_{\text{sample}} \mathbf{P} \cdot \mathbf{E}_0^* d\tau; \qquad \mathbf{P} = \chi_e \mathbf{E}$$

where $\mathbf{E}_0$ and $\mathbf{H}_0$ are the electric and magnetic fields in the empty cavity, $\mathbf{E}$ and $\mathbf{H}$ are the corresponding quantities in the perturbed cavity; $\mathbf{M}$ and $\mathbf{P}$ are the magnetic and electric polarizations in the sample, and χ and χ_e are the magnetic and electric susceptibilities of the sample.

In any particular geometrical and modal situation, the right side of (1) must be evaluated and the result is then a relation between $(d\omega/\omega_0)$ on the one hand and μ, κ and cavity and sample geometry on the other. The assumption in the perturbation theory is that electric and magnetic fields just outside the sample are their (known) empty cavity values. From this and the requirement of continuity of tangential E and H and normal B and D at air-sample interfaces, we can obtain the fields inside the sample and so calculate the numerator of (1). This is the approach used by the authors in Refs. 1 and 2 and we shall continue to follow this.

Instead of specializing to a particular cavity operating in a particular mode, as is done by the various authors in Ref. 1, we feel it would be quite useful to assume the ferrite sample placed in an axial cavity of generalized cross section operating in any TE, TM or TEM mode and find the frequency shift produced by this sample. We shall assume: the cavity has a z axis which coincides with the z axis of the sample; the sample is a disc of arbitrary cross section; the sample volume is small compared to the cavity volume; the sample is magnetized along

the z axis; the cavity has perfectly conducting walls. The result of our endeavor is to reduce the right side of (1) to an expression containing a series of contour integrals only, in which μ and κ are explicit coefficients of these integrals. The contour integrals are integrals of empty cavity fields (or potentials generating these fields) taken around the contour of the cavity and the contour of the sample in the transverse plane. For any particular geometry and modal distribution, these contour integrals are easily evaluated, and we give examples of their evaluation in the circular coaxial cavity, circular cylinder cavity and rectangular waveguide cavity for the TE_{pqN}, TM_{pqN}, TM_{pq0} modes. In these examples, we consider thin ferrite discs. Thus the frequency shift, within the confines of a perturbation theory, is obtained for a quite generalized cavity operating in any of its natural modes of oscillation; the frequency shift for any given cavity operating in any given mode can then be obtained from the general result.

An interesting fact that emerges from the general result is that the expression for $(d\omega/\omega_0)$ is independent of κ whenever the explicit time independent part of the field, or potential, is real. This reality of the potential corresponds to a linear polarization at any point in the transverse plane and since a linear polarization is equivalent to two equal and opposite circular polarizations (corresponding to $+\kappa$ and $-\kappa$ values of the off-diagonal element) the net effect is for κ to cancel out of the interaction of the linear wave with the ferrite at each point of the ferrite. This is the situation in the rectangular waveguide cavity with a rectangular slab sample, as we shall see later (Section 2.2.6). It is also the case in the TEM-type cavity (Section 2.2.5) since here, too, the fields are real and therefore linearly polarized at any given point in the ferrite. These cases are thus not suited for determining the off-diagonal component of the Polder tensor.

On the other hand, we shall see that when we deal with a circular coaxial cavity or circular cylinder cavity (Sections 2.2.1 to 2.2.4) and choose a circularly polarized mode, the time independent part of the potential or field is complex and goes as $e^{ip\theta}$ where θ is angle and p is angular mode number. In this case the circular polarization $e^{i(p\theta+\omega t)}$ interacts in an unbalanced way with the precessing spins and leads to a coupling to κ. The expression for $d\omega/\omega_0$ thus involves both μ and κ and is ideally suited for use in experimental determinations of Polder tensor elements.

The existence of complex potentials is associated with a degenerate state of the system. Since degeneracy is usually related to a symmetric structure (e.g. square or circular guide), we state that in almost all

cases the only structures suitable for measuring κ in thin ferrite samples are symmetric ones. There are, however, asymmetric structures which also have degenerate states.

II. ANALYSIS

2.1 *Frequency Shift Produced by Thin Ferrite Disc Sample in Cavity of Generalized Cross Section.*

2.1.1 TE_{pqN} *and* TM_{pqN} *type modes.* We consider a cavity of arbitrary cross section with an axis in the z direction containing a thin disc of ferrite of small cross section. The ferrite sample is magnetized in the z direction (see Fig. 1). We assume the cavity walls are perfect, lossless conductors. We shall consider standing waves in this cavity, which are the result of superposing the traveling waves $\pm f(u, v)e^{\pm j\beta z}$, where u, v are transverse coordinates and β is the longitudinal propagation number.

The electric field in the empty cavity which satisfies the boundary condition that transverse $\mathbf{E}_0$ vanish at $z = 0$ and the condition div $\mathbf{E}_0 = 0$, is obtained as a superposition of $e^{\pm j\beta z}$ waves and is given by

$$\mathbf{E}_0 = \mathbf{E}_t \sin \beta z + E_z \cos \beta z \, \mathbf{z}_0 \tag{3}$$

where $\mathbf{E}_t$ and $\mathbf{E}_z$ are functions of transverse coordinates only. The condi-
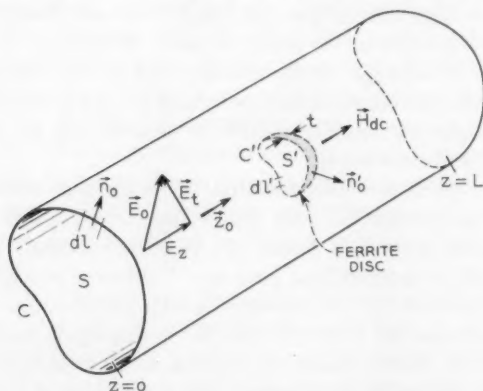

Fig. 1 — Geometry of cavity and sample. C, C', S and S' are contours and cross sections of transverse section of cavity and ferrite sample, respectively; t is thickness of ferrite disc; $\mathbf{z}_0$ is unit vector along axis of cavity; $\mathbf{H}_{dc}$ is the steady applied magnetic field; $\mathbf{E}_t$ is the transverse component of electric field $\mathbf{E}_0$; $\mathbf{n}_0$ and $\mathbf{n}_0'$ are outward unit normals to C and C' and L is the length of the cavity.

tion that transverse $\mathbf{E}_0$ vanish at $z = L$ leads to

$$\beta L = N\pi, \qquad N = 1, 2, \cdots, \quad (4)$$

where N is the longitudinal mode number.

We shall attempt to write the transverse electric field in terms of two potentials φ and ψ which are functions of the transverse coordinates:

$$\mathbf{E}_t = \nabla_t\varphi + \mathbf{z}_0 \times \nabla_t\psi \qquad (5)$$

where ∇_t is the transverse gradient operator. With $\nabla = \nabla_t + \mathbf{z}_0(\partial/\partial z)$ and $\nabla \cdot \mathbf{E}_0 = 0$, we find from (3) and (5)

$$\nabla_t^2\varphi - \beta E_z = 0. \qquad (6)$$

Since the z component of $\mathbf{E}_0$, $E_z \cos \beta z$, satisfies the wave equation, we find with (6) that φ satisfies

$$
\begin{aligned}
(\nabla_t^2 + k_c^2)\varphi &= 0, \qquad E_z = -\frac{k_c^2}{\beta}\varphi, \\
k_c^2 &= k^2 - \beta^2, \\
k &= \frac{2\pi}{\lambda_0} = \omega\sqrt{\mu_0\epsilon_0}, \\
\beta &= \frac{2\pi}{\lambda_g},
\end{aligned}
\qquad (7)
$$

where λ_0 and λ_g are free space wavelength and cavity wavelength, respectively.

In like manner, the transverse component of $\mathbf{E}_0$, namely

$$(\nabla_t\varphi + \mathbf{z}_0 \times \nabla_t\psi) \sin \beta z,$$

satisfies the wave equation and this leads to

$$(\nabla_t^2 + k_c^2)\psi = 0. \qquad (8)$$

The complete electric and magnetic fields in the empty cavity are thus given by

$$
\begin{aligned}
\mathbf{E}_0 &= (\nabla_t\varphi + \mathbf{z}_0 \times \nabla_t\psi) \sin \beta z - \frac{k_c^2}{\beta}\varphi \cos \beta z \mathbf{z}_0, \\
\mathbf{H}_0 &= \frac{j}{\omega\mu_0}\left(\nabla_t + \mathbf{z}_0 \frac{\partial}{\partial z}\right) \times \mathbf{E}_0 \\
&= \frac{j\beta}{\omega\mu_0}\left[-\nabla_t\psi \cos \beta z + \frac{k^2}{\beta^2}\mathbf{z}_0 \times \nabla_t\varphi \cos \beta z - \frac{kc^2}{\beta}\psi \sin \beta z \mathbf{z}_0 \right].
\end{aligned}
\qquad (9)
$$

The existence of a φ wave alone can be interpreted as a TM wave as (9) shows. The boundary condition on φ is determined by $E_z = 0$ on C, and from (7) this is equivalent to $\varphi_c = 0$. If $\varphi_c = 0$, then inspection of (9) shows that H_{normal} also vanishes on C. Transverse $\mathbf{E}$ is zero on the end plates of the cavity provided $\beta L = N\pi$.

The existence of a ψ wave alone can be interpreted as a TE wave. The boundary condition that $H_{\text{normal}} = 0$ on all conducting surfaces is satisfied if $\beta L = N\pi$ and if $(\partial \psi/\partial n) = 0$ on C, where $\mathbf{n}$ is the unit normal directed outward from C. If $(\partial \psi/\partial n) = 0$ on C, we see from (9) that tangential $\mathbf{E}$ is also zero on C, as required.

The problem now is to find contour integral expressions for the quantities $W_s^{(M)}$, $W_s^{(E)}$ and $W_c^{(E)}$ defined in (1) and (2). For this purpose, we shall make use of the following three relations:

$$\int_S |\nabla_t \xi|^2 \, dS = k_c^2 \int_S |\xi|^2 \, dS + \oint_C (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0 \, dl, \tag{10}$$

$$2k_c^2 \int_S |\xi|^2 \, dS = \oint_C [(\nabla_{k_c} \xi \cdot \mathbf{k}_c)(\nabla_t \xi^* \cdot \mathbf{n}_0) - \xi^* \nabla_t (\nabla_{k_c} \xi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0] \, dl, \tag{11}$$

$$\int_S \nabla_t \xi \cdot (\mathbf{z}_0 \times \nabla_t \xi^*) \, dS = \oint_C (\xi \nabla_t \xi^*) \cdot d\mathbf{l}. \tag{12}$$

Here ∇_{k_c} is the gradient operator in $\mathbf{k}_c$ space and ξ stands for either φ or ψ. Equations (10), (11) and (12) apply to either C and S or to C' and S'. The above relations are derived in Appendices I, II and III, respectively.

Calculating first the electric energy stored in the cavity we have

$$W_c^{(E)} = \frac{\epsilon_0}{2} \int \mathbf{E}_0 \cdot \mathbf{E}_0^* \, d\tau = \frac{\epsilon_0 L}{4} \int_S \left[|\nabla_t \varphi|^2 + |\nabla_t \psi|^2 \right.$$
$$\left. + \left(\frac{k_c^2}{\beta}\right) |\varphi|^2 \right] dS, \tag{13}$$

where we are assuming that a φ and ψ wave do not exist simultaneously. Making use of (10) and (11) with the condition that $\varphi = 0$ on C and $(\partial \psi^*/\partial n) = 0$ on C, we find

$$W_c^{(E)} = \frac{\epsilon_0 L}{8} \frac{k^2}{\beta^2} \oint_C (\nabla_{k_c} \varphi \cdot \mathbf{k}_c)(\nabla_t \varphi^* \cdot \mathbf{n}_0) \, dl \qquad \text{for } \varphi \text{ waves;}$$
$$\tag{14}$$
$$W_c^{(E)} = -\frac{\epsilon_0 L}{8} \oint_C \psi^* \nabla_t (\nabla_{k_c} \psi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0 \, dl \qquad \text{for } \psi \text{ waves.}$$

We next calculate the magnetic energy stored in the sample:

$$W_s^{(M)} = \frac{\mu_0}{2} \int \mathbf{M} \cdot \mathbf{H}_0^* d\tau',$$

$$\mathbf{M} = \chi \mathbf{H}; \qquad \chi = \begin{pmatrix} \mu - 1 & j\kappa & 0 \\ -j\kappa & \mu - 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{15}$$

According to the perturbation theory, the $\mathbf{H}_0$ and $\mathbf{H}$ appearing in (15) are the RF magnetic fields in the empty cavity and ferrite sample, respectively. From (15), $M_z = 0$; therefore $(H_z)_0^*$ is not needed in the expression for $W_s^{(M)}$. Continuity of tangential H from air to sample shows that we may use $\mathbf{H}_0$ in place of $\mathbf{H}$ in (15). The field to be used throughout (15) is, therefore, that given by (9) without the $(H_z)_0$ component:

$$\mathbf{H}_0 = \frac{j\beta}{\omega\mu_0} \left[-\nabla_t \psi \cos \beta z + \frac{k^2}{\beta^2} \mathbf{z}_0 \times \nabla_t \varphi \cos \beta z \right]. \tag{16}$$

When (16) is used in (15) we find (Appendix IV)

$$W_s^{(M)} = \frac{f\beta^2 \cos^2 \beta z_0}{2\omega^2 \mu_0} t \int_{s'} [(\mu - 1) \mid \nabla_t \xi \mid^2 + j\kappa \nabla_t \xi \cdot \mathbf{z}_0 \times \nabla_t \xi^*] dS' \tag{17}$$

where z_0 (not to be confused with $\mathbf{z}_0$) is the position of the sample along the axis of the cavity and $f = 1$ or (k^2/β^2) for $\xi = \psi$ and $\xi = \varphi$, respectively. It is assumed $t \ll L$. Transforming (17) to contour integrals through (10), (11) and (12), we have

$$W_s^{(M)} = \frac{f\beta^2 \cos^2 \beta z_0}{2\omega^2 \mu_0} t \left\{ (\mu - 1) \oint_{c'} (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0' \, dl' \right.$$

$$+ \frac{\mu - 1}{2} \oint_{c'} [(\nabla_{k_c} \xi \cdot \mathbf{k}_c)(\nabla_t \xi^* \cdot \mathbf{n}_0') - \xi^* \nabla_t (\nabla_{k_c} \xi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0'] \, dl' \tag{18}$$

$$+ j\kappa \oint_{c'} (\xi \nabla_t \xi^*) \cdot d\mathbf{l}' \left. \right\}.$$

The final step is the calculation of the electric energy stored in the sample:

$$W_s^{(E)} = \frac{\epsilon_0}{2} \int \mathbf{P} \cdot \mathbf{E}_0^* \, d\tau^h, \qquad \mathbf{P} = \chi_e \mathbf{E}, \qquad \chi_e = \epsilon - 1, \tag{19}$$

ϵ being the dielectric constant of the sample and $\mathbf{E}_0$ and $\mathbf{E}$ the electric fields in empty cavity and inside ferrite, respectively. Since $\mathbf{E}_t = \mathbf{E}_{t_0}$

and $\epsilon_0 E_{z_0} = \epsilon_0 \epsilon E_z$, we have

$$
\begin{aligned}
W_s{}^{(E)} &= \frac{\epsilon_0(\epsilon - 1)}{2} \int \left[\mid E_{t_0} \mid^2 + \frac{1}{\epsilon} \mid E_{z_0} \mid^2 \right] d\tau' \\
&= \frac{\epsilon_0(\epsilon - 1)t}{2} \left[\sin^2 \beta z_0 \int_{s'} (\mid \nabla_t \varphi \mid^2 + \mid \nabla_t \psi \mid^2) \, dS' \right. \\
&\qquad\qquad \left. + \frac{\cos^2 \beta z_0}{\epsilon} \left(\frac{k_c{}^2}{\beta} \right)^2 \int_{s'} \mid \varphi \mid^2 dS' \right].
\end{aligned}
$$
(20)

Making use of (10) and (11), we find

$$
\begin{aligned}
W_s{}^{(E)} &= \frac{\epsilon_0(\epsilon - 1)}{2} \, t \left\{ \sin^2 \beta z_0 \left[\oint_{c'} (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0' \, dl' \right.\right. \\
&\qquad \left. + \frac{1}{2} \oint_{c'} [(\nabla_{k_c} \xi \cdot \mathbf{k}_c)(\nabla_t \xi^* \cdot \mathbf{n}_0') - \xi^* \nabla_t (\nabla_{k_c} \xi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0'] \, dl' \right] \\
&\quad \left. + \frac{\cos^2 \beta z_0}{2\epsilon} \left(\frac{k_c}{\beta} \right)^2 \oint_{c'} [(\nabla_{k_c} \xi \cdot \mathbf{k}_c)(\nabla_t \xi^* \cdot \mathbf{n}_0') - \xi^* \nabla_t (\nabla_{k_c} \xi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0'] \, dl' \right\}
\end{aligned}
$$
(21)

where for TM modes (φ waves) both the $\sin^2 \beta z_0$ and $\cos^2 \beta z_0$ terms are needed, and for TE waves (ψ waves) only the $\sin^2 \beta z_0$ term is needed.

2.1.2 *TM_{pq0} type modes.* The TM_{pq0} type modes are obtained by considering $\beta = 0$ in all $e^{\pm j\beta z}$ field dependences, i.e. the cavity is cut off in the z direction and there is to be no field dependence on z. In this case we have

$$
\begin{aligned}
(\nabla_t{}^2 + k^2)E_z &= 0, \\
k_c{}^2 = k^2 - \beta^2 = k^2, &\qquad \mathbf{k}_c = \mathbf{k}.
\end{aligned}
$$
(22)

Since transverse $\mathbf{E}$ must be zero at $z = 0$ and $z = L$, and since there can be no z dependence in this case, transverse E must vanish everywhere in the cavity. There can only be an E_z component. Thus a TE_{pq0} mode can not exist, for then $\mathbf{E}_0 \equiv 0$ everywhere and $\mathbf{H}_0 \propto \text{curl } \mathbf{E}_0$ gives $\mathbf{H}_0 \equiv 0$ also. A TM_{pq0} mode is the only possibility and we have

$$
\begin{aligned}
\mathbf{E}_0 &= E_z \mathbf{z}_0, \\
\mathbf{H}_0 &= \frac{j}{\omega \mu_0} \nabla \times \mathbf{E}_0 = \frac{-j}{\omega \mu_0} \mathbf{z}_0 \times \nabla_t E_z,
\end{aligned}
$$
(23)

where E_z is a function of transverse coordinates only. E_z plays the same role now as φ did in the TM_{pqN} case before. The boundary condition $E_z = 0$ on C causes H_{normal} to vanish automatically on C, as required, as is seen from (23).

Proceeding exactly as in the steps leading to (14), (18) and (21), we find

$$W_c^{(E)} = \frac{\epsilon_0 L}{4k^2} \oint_C (\nabla_k E_z \cdot \mathbf{k})(\nabla_l E_z^* \cdot \mathbf{n}_0) \, dl, \tag{24}$$

$$W_s^{(M)} = \frac{t}{2\omega^2\mu_0} \left\{ (\mu - 1) \oint_{C'} (E_z \nabla_l E_z^*) \cdot \mathbf{n}_0' \, dl' \right.$$

$$+ \frac{\mu - 1}{2} \oint_{C'} [(\nabla_k E_z \cdot \mathbf{k})(\nabla_l E_z^* \cdot \mathbf{n}_0') - E_z^* \nabla_l (\nabla_k E_z \cdot \mathbf{k}) \cdot \mathbf{n}_0'] \, dl' \tag{25}$$

$$\left. + j\kappa \oint_{C'} (E_z \nabla_l E_z^*) \cdot d\mathbf{l}' \right\},$$

$$W_s^{(E)} = \frac{\epsilon_0(\epsilon - 1)t}{4\epsilon k^2} \oint_{C'} [(\nabla_k E_z \cdot \mathbf{k})(\nabla_l E_z^* \cdot \mathbf{n}_0')$$

$$- E_z^* \nabla_l (\nabla_k E_z \cdot \mathbf{k}) \cdot \mathbf{n}_0'] \, dl'. \tag{26}$$

2.1.3 *TEM type cavity.* In this case $\lambda_g = \lambda_0$, $\beta = k$, and $k_c^2 = 0$. We can write the fields in terms of a single potential φ:

$$\mathbf{E}_0 = (\nabla_t \varphi) \sin \beta z,$$

$$\mathbf{H}_0 = \frac{j}{\omega\mu_0} \nabla \times \mathbf{E}_0 = \frac{j\beta}{\omega\mu_0} (\mathbf{z}_0 \times \nabla_t \varphi) \cos \beta z, \tag{27}$$

where φ satisfies div $\mathbf{E}_0 = 0$ or

$$\nabla_t^2 \varphi = 0. \tag{28}$$

The boundary condition $E_{\text{tangential}} = 0$ on cavity walls gives $(\partial\varphi/\partial l) = 0$ or $\varphi = $ constant on C. This automatically makes $H_{\text{normal}} = 0$ on cavity walls as we see from (27). The end plate condition on transverse $\mathbf{E}$ again gives $\beta L = N\pi$.

The various stored energies are found as before, with use made of (10), (11) and (12):

$$W_c^{(E)} = \frac{\epsilon_0}{2} \int \mathbf{E}_0 \cdot \mathbf{E}_0^* \, d\tau = \frac{\epsilon_0 L}{4} \int_S |\nabla_t \varphi|^2 \, dS$$

$$= \frac{\epsilon_0 L}{4} \oint_C (\varphi \nabla_t \varphi^*) \cdot \mathbf{n}_0 \, dl, \tag{29}$$

$$W_s^{(M)} = \frac{\beta^2 \cos^2 \beta z_0}{2\omega^2\mu_0} t \left[(\mu - 1) \oint_{C'} (\varphi \nabla_t \varphi^*) \cdot \mathbf{n}_0' \, dl' \right.$$

$$\left. + j\kappa \oint_{C'} (\varphi \nabla_t \varphi^*) \cdot d\mathbf{l}' \right], \tag{30}$$

$$W_s^{(E)} = \frac{\epsilon_0}{2} \int \mathbf{P} \cdot \mathbf{E}_0{}^* \, d\tau' = \frac{\epsilon_0(\epsilon - 1)}{2} \int | \nabla_t \varphi |^2 \sin^2 \beta z' \, dS' \, dz'$$

$$= \frac{\epsilon_0(\epsilon - 1)}{2} t \sin^2 \beta z_0 \int_{S'} | \nabla_t \varphi |^2 \, dS' \qquad (31)$$

$$= \frac{\epsilon_0(\epsilon - 1)}{2} t \sin^2 \beta z_0 \oint_{c'} (\varphi \nabla_t \varphi^*) \cdot \mathbf{n}_0' \, dl',$$

where we have, in (31), used the fact that $\mathbf{E}$ is purely transverse and must be continuous from air to sample ($\mathbf{E} = \mathbf{E}_0$).

2.2 *Circular Coaxial Cavity (Higher Modes); Circular Cylindrical Cavity; TEM Type Cavity; Rectangular Cavity.*

We are now in a position to calculate the frequency shift in various cases from (1).

2.2.1 *Circular coaxial cavity, ferrite ring, TE_{pqN} modes.* The geometry we are considering is shown in Fig. 2.

We are here dealing with a ψ wave only. Consider the sample placed at $\sin \beta z_0 = 0$ (node of the electric field) so that from (21) there is no electric energy stored in the sample.

The solution of $(\nabla_t{}^2 + k_c{}^2)\psi = 0$ for the empty cavity potential in this geometry is

$$\psi = e^{jp\theta}[J_p(k_c r) + CN_p(k_c r)],$$
$$= e^{jp\theta}Z_p(k_c r), \qquad a_1 < r < a_2, \qquad (32)$$

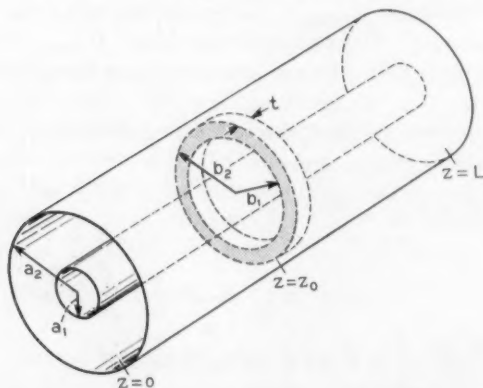where J_p and N_p are Bessel functions of the first and second kind of

Fig. 2 — Circular coaxial cavity with ferrite ring of thickness t placed at $z = z_0$; a_1 and a_2 are inner and outer radii of cylinders; b_1 and b_2 are inner and outer radii of ferrite ring.

order p, Z_p is the general cylinder function and C is an arbitrary constant to be determined by the boundary conditions. Z_p satisfies[3]

$$Z_p''(x) + \frac{1}{x} Z_p'(x) + \left(1 - \frac{p^2}{x^2}\right) Z_p(x) = 0,$$

$$Z_p' = Z_{p-1} - \frac{p}{x} Z_p .$$

$$\tag{33}$$

The boundary condition $(\partial\psi/\partial n) = 0$ on C $(r = a_1, a_2)$ gives

$$Z_p'(k_c a_1) = Z_p'(k_c a_2) = 0,$$

$$-C = \frac{J_p'(k_c a_1)}{N_p'(k_c a_1)} = \frac{J_p'(k_c a_2)}{N_p'(k_c a_2)},$$

$$k_c a_2 = \sigma_{pq}, \qquad p, q = \text{integers},$$

$$\sigma_{pq} = q\text{th root of } J_p'(k_c a_1)N_p'(k_c a_2) - J_p'(k_c a_2)N_p'(k_c a_1) = 0,$$

$$\beta L = N\pi.$$

$$\tag{34}$$

When (32) and (33) are used in (14) and (18) (remembering that C consists of the contours $r = a_1$ and $r = a_2$ taken in opposite senses and C' consists of the contours $r = b_1$, $r = b_2$ also taken oppositely) we find

$$\frac{d\omega}{\omega_0} =$$

$$\frac{\frac{\beta^2}{k^2} \frac{t}{L} \left[(\mu - 1) \left\{ Z_p^2 \left(1 - \frac{2p}{k_c^2 b^2}\right) + \frac{2}{k_c b} Z_p Z_{p-1}(1 - p) \right.\right.}{\qquad} $$

$$\left.\left. + Z_{p-1}^2 \right\} \pm \frac{2\kappa p}{k_c^2 b^2} Z_p^2 \right] k_c^2 b^2 \Big\}_{b_1}^{b_2} $$

$$\overline{\left\{ k_c^2 a^2 \left(1 - \frac{p^2}{k_c^2 a^2}\right) Z_p^2(k_c a) \right\}_{a_1}^{a_2}}$$

$$\tag{35}$$

with $(\beta^2/k^2) = (\lambda_0/\lambda_g)^2$, $\beta L = N\pi$, $k_c a_2 = \sigma_{pq} = $ roots of (34). The $\pm\kappa$ term arises from each of the two types of circular polarizations which are possible, i.e. $e^{\pm jp\theta}$. The functions Z_p in the numerator are defined by $Z_p = Z_p(k_c b)$.

2.2.2 *Circular coaxial cavity, ferrite ring, TM_{pqN} modes.* Here we are considering a φ wave with the sample again placed at $\sin \beta z_0 = 0$. The solution of $(\nabla_t^2 + k_c^2)\varphi = 0$ in the geometry of Fig. 2 is again given by (32), but the boundary condition $\varphi = 0$ on $r = a_1, a_2$ now gives

$$Z_p(k_c a_1) = Z_p(k_c a_2) = 0,$$

$$-C = \frac{J_p(k_c a_1)}{N_p(k_c a_1)} = \frac{J_p(k_c a_2)}{N_p(k_c a_2)},$$

$$k_c a_2 = \tau_{pq}, \qquad p, q = \text{integers},$$

$$\tau_{pq} = q\text{th root of } J_p(k_c a_1)N_p(k_c a_2) - J_p(k_c a_2)N_p(k_c a_1) = 0,$$

$$\beta L = N\pi.$$

$$\tag{36}$$

The evaluation of the cavity energy, magnetic energy in a sample, and electric energy in a sample [(14), (18) and (21)] is straightforward and leads to the following result for the frequency shift:

$$\frac{d\omega}{\omega_0} =$$

$$\frac{t}{L}\left[(\mu - 1)\left\{ Z_p^2\left(1 - \frac{2p}{k_c^2 b^2}\right) + \frac{2}{k_c b}Z_p Z_{p-1}(1-p) + Z_{p-1}^2 \right\} \right. \tag{37}$$

$$\left. \pm \frac{2\kappa p}{k_c^2 b^2} Z_p^2 + \frac{\epsilon - 1}{\epsilon}\frac{k_c^2}{k^2}\left\{ Z_p^2 - \frac{2p}{k_c b}Z_p Z_{p-1} + Z_{p-1}^2 \right\}\right] k_c^2 b^2 \Big\}_{b_1}^{b_2}$$
$$\overline{\left\{ k_c^2 a^2 Z_{p-1}^2(k_c a)\right\}_{a_1}^{a_2}}$$

with $(k_c^2/k^2) = 1 - (\lambda_0/\lambda_g)^2$, $\lambda_g = (2L/N)$, $k_c a_2 = \tau_{pq} = $ roots of (36). The argument of the cylinder functions in the numerator of (37) is implicitly $k_c b$.

2.2.3 *Circular coaxial cavity, ferrite ring, TM_{pq0} modes.* The solution of $(\nabla_t^2 + k^2)E_z = 0$ in this geometry is

$$E_z = e^{jp\theta}Z_p(kr),$$
$$Z_p = J_p + CN_p. \tag{38}$$

The boundary condition $E_z = 0$ on $r = a_1$, a_2 gives the same set of equations as in (36) with k_c replaced by k, except that $\beta L = N\pi$ is not now applicable. When (38), (36) and (33) are used in (24), (25) and (26) we find for the frequency shift

$$\frac{d\omega}{\omega_0} =$$

$$\frac{t}{L}\left[(\mu - 1)\left\{ Z_p^2\left(1 - \frac{2p}{k^2 b^2}\right) + \frac{2}{kb}Z_p Z_{p-1}(1-p) + Z_{p-1}^2 \right\} \right. \tag{39}$$

$$\left. \pm \frac{2\kappa p}{k^2 b^2} Z_p^2 + \frac{\epsilon - 1}{\epsilon}\left\{ Z_p^2 - \frac{2p}{kb}Z_p Z_{p-1} + Z_{p-1}^2 \right\}\right] b^2 \Big\}_{b_1}^{b_2}$$
$$\overline{2\left\{ a^2 Z_{p-1}^2(ka)\right\}_{a_1}^{a_2}}$$

with $ka_2 = \tau_{pq}$, $\tau_{pq} = $ roots of (36). The numerator argument is kb.

2.2.4 *Circular cylindrical cavity, circular ferrite disc, TE_{pqN}, TM_{pqN}, TM_{pq0} modes.* The geometry of this situation is obtained from Fig. 2 by letting a_1, b_1 approach 0, so that the ferrite ring becomes a disc. The frequency shifts in the various modal cases can then be obtained from (35), (37) and (39) by letting a_1, b_1 approach 0 and putting $C = 0$ in (32), (34), (36) and (38), since only the Bessel function which is regular at

$r = 0$ (J_p) is now allowed. Under these conditions the following quantities approach zero for $p \geqq 0$:

$$r^2 \left(1 - \frac{2p}{r^2}\right) J_p^{\,2}(r), \qquad 2rJ_p(r)J_{p-1}(r)(1 - p),$$

$$r^2 J_{p-1}^{\,2}(r), \qquad pJ_p^{\,2}(r), \qquad r^2 J_p^{\,2}(r), \qquad 2prJ_p(r)J_{p-1}(r)'$$

where $r = k_c a_1$ or $k_c b_1$ approaches 0. In other words, all quantities at the lower limits vanish in (35), (37) and (39) for this geometry. Equation (35) becomes for TE_{pqN} modes:

$$\frac{d\omega}{\omega_0} =$$

$$\frac{\beta^2}{k^2} \frac{t}{L} \frac{b_2^{\,2}}{a_2^{\,2}} \left[(\mu - 1) \left\{ J_p^{\,2} \left(1 - \frac{2p}{k_c^2 b^2}\right) \right.\right.$$
$$\left.\left. + \frac{2}{k_c b} J_p J_{p-1}(1 - p) + J_{p-1}^{\,2} \right\} \pm \frac{2\kappa p}{k_c^2 b^2} \right]_{b=b_2}$$
$$\frac{}{\left[\left(1 - \frac{p^2}{k_c^2 a^2}\right) J_p^{\,2}(k_c a) \right]_{a=a_2}} \tag{40}$$

with $(\beta^2/k^2) = (\lambda_0/\lambda_g)^2$, $\beta L = N\pi$, $k_c a_2 = s_{pq} =$ roots of $J_p{}'(k_c a_2) = 0$, $p, q =$ integers. The numerator argument is $k_c b$.

For TM_{pqN} modes, (37) becomes

$$\frac{d\omega}{\omega_0} =$$

$$\frac{t}{L} \frac{b_2^{\,2}}{a_2^{\,2}} \left[(\mu - 1) \left\{ J_p^{\,2} \left(1 - \frac{2p}{k_c^2 b^2}\right) + \frac{2}{k_c b} J_p J_{p-1}(1 - p) + J_{p-1}^{\,2} \right\}\right.$$

$$\left.\pm \frac{2\kappa p}{k_c^2 b^2} J_p^{\,2} + \frac{(\epsilon - 1)}{\epsilon} \frac{k_c^{\,2}}{k^2} \left\{ J_p^{\,2} - \frac{2p}{k_c b} J_p J_{p-1} + J_{p-1}^{\,2} \right\} \right]_{b=b_2}$$
$$\frac{}{J_{p-1}^{\,2}(k_c a_2)} \tag{41}$$

$$\frac{k_c^{\,2}}{k^2} = 1 - \left(\frac{\lambda_0}{\lambda_g}\right)^2, \qquad \lambda_g = \frac{2L}{N},$$

with $k_c a_2 = t_{pq} =$ roots of $J_p(k_c a_2) = 0$, $p, q =$ integers. The numerator argument is $k_c b$.

For TM_{pq0} modes, (39) becomes

$$\frac{d\omega}{\omega_0} =$$

$$\frac{t}{L}\frac{b_2^2}{a_2^2}\left[(\mu-1)\left\{J_p^2\left(1-\frac{2p}{k^2b^2}\right)+\frac{2}{kb}J_pJ_{p-1}(1-p)+J_{p-1}^2\right\}\right.$$

$$\underline{\left.\pm\frac{2\kappa p}{k^2b^2}J_p^2+\frac{\epsilon-1}{\epsilon}\left\{J_p^2-\frac{2p}{kb}J_pJ_{p-1}+J_{p-1}^2\right\}\right]_{b=b_2}}$$

$$\overline{2J_{p-1}^2(ka_2)}$$

(42)

$ka_2 = t_{pq} = $ roots of $J_p(ka_2) = 0$, $k = (2\pi/\lambda_0)$, p, $q = $ integers. Again the sample is considered placed at $\sin \beta z_0 = 0$ in (40) and (41). The numerator argument is kb.

2.2.5 *TEM type cavity.* In this case the quantity φ appearing in (29), (30) and (31) is real so that

$$\oint_{c'}\varphi\nabla_t\varphi^*\cdot\mathbf{dl'} = \frac{1}{2}\oint_{c'}\nabla_t\varphi^2\cdot\mathbf{dl'} = \frac{1}{2}\oint_{c'}\frac{\partial(\varphi^2)}{\partial l'}\,dl' \equiv 0.$$

Consequently, there is no coupling to κ and this type of cavity is not suited for measurement of the off-diagonal component of the Polder tensor. If the sample is placed at $\sin \beta z_0 = 0$, the frequency shift is given by

$$\frac{d\omega}{\omega_0} = \frac{t}{L}(\mu-1)\frac{\displaystyle\oint_{c'}\frac{\partial(\varphi^2)}{\partial n'}\,dl'}{\displaystyle\oint_{c}\frac{\partial(\varphi^2)}{\partial n}\,dl}$$

(43)

where $\partial/\partial n'$ is the derivative along the normal to C or C'.

2.2.6 *Rectangular cavity, thin rectangular ferrite slab, TE_{pqN}, TM_{pqN} modes.* The geometry is shown in Fig. 3.
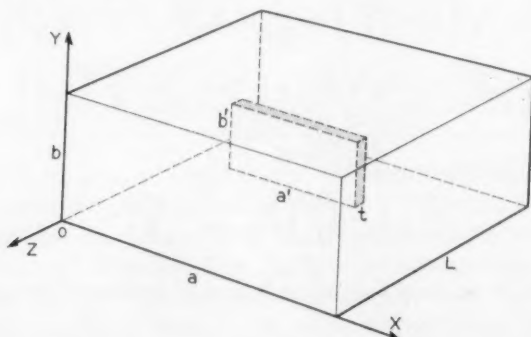

Fig. 3 — Rectangular cavity containing rectangular ferrite slab.

For TE_{pqN} modes, $\psi = \cos k_x x \cos k_y y$. The boundary condition $\partial \psi / \partial n = 0$ on C $(x = 0, x = a, y = 0, y = b)$ gives $k_x a = p\pi, k_y b = q\pi$ where p, q = integers. The wave equation $(\nabla_t^2 + k_c^2)\psi = 0$ then gives $k_c^2 = k_x^2 + k_y^2 = k^2 - \beta^2$. Again $\beta L = N\pi$.

For TM_{pqN} modes, $\varphi = \sin k_x x \sin k_y y$ and this satisfies the boundary condition $\varphi = 0$ on C if $k_x a = p\pi, k_y b = q\pi$. Again $k_c^2 = k_x^2 + k_y^2$ and $\beta L = N\pi$.

In each of these cases it is clear from (32) that the ferrite sample does not couple to κ since $\xi = \varphi$ or ψ is real. As a result,

$$\kappa \oint_{c'} \xi \nabla_t \xi^* \cdot \mathbf{dl}' = \frac{\kappa}{2} \oint_{c'} \nabla_t (\xi^2) \cdot \mathbf{dl}' = 0,$$

unlike the situation in the circular coaxial cavity [(35), (37) and (39)]. Thus the situation described here, as with the TEM cavity, is not suited to a measurement of κ. We shall, however, write down the frequency shift in the TE_{pqN} case.

We shall again consider the sample placed at $\sin \beta z_0 = 0$, so that there is no electric energy stored in the sample (21). After evaluating the various quantities needed in (14) and (18) and carrying out the integrations, we find for the frequency shift of a transversely centered slab.

$$\frac{d\omega}{\omega_0} = \frac{t}{L} \frac{\beta^2}{k^2} \frac{a'b'}{ab} (\mu - 1) \left[1 + (-1)^{p+q+1} \frac{\sin \theta_1}{\theta_1} \frac{\sin \theta_2}{\theta_2} \right.$$
$$\left. + (-1)^p \frac{k_y^2 - k_x^2}{k_y^2 + k_x^2} \frac{\sin \theta_1}{\theta_1} + (-1)^q \frac{k_x^2 - k_y^2}{k_x^2 + k_y^2} \frac{\sin \theta_2}{\theta_2} \right], \quad (44)$$

$$\frac{\beta^2}{k^2} = \left(\frac{\lambda_0}{\lambda_g} \right)^2, \quad \beta L = N\pi, \quad k_x = \frac{p\pi}{a}, \quad k_y = \frac{q\pi}{b}, \quad \theta_1 = k_x a', \quad \theta_2 = k_y b',$$

p, q = integers.

III. REMARKS

In the case of TM_{pqN} modes in the circular coaxial cavity or circular cylindrical cavity, we see from (37) and (41) that the frequency shift depends on the values of the three quantities μ, $|\kappa|$ and ϵ. In these cases we placed the sample at $\sin \beta z_0 = 0$. If now we place the sample at $\cos \beta z_0 = 0$, (18) shows that the stored magnetic energy in the sample would be zero while (21) would give a stored electric energy in the sample proportional to $\epsilon - 1$. The frequency shift corresponding to (37) and (41) for this situation would thus depend only on ϵ. From these two situations we could infer all three quantities μ, κ and ϵ. A similar argument holds in the TE_{pqN} case in circular coaxial or circular cylindrical

cavities [(35) and (40)]. If we placed the sample at $\cos \beta z_0 = 0$ we would again find zero magnetic energy in the sample and now $d\omega/\omega_0$ would depend only on $\epsilon - 1$ [see (18) and (21)]. Thus, we could again infer the values of μ, κ and ϵ separately. In the TM_{pq0} case, on the other hand, we can never separate the effects of μ and κ on the one hand and ϵ on the other, since the stored magnetic and electric energies in the sample are independent of longitudinal position of sample, as (25) and (26) show.

An interesting mathematical point is the manner in which μ, κ and ϵ appear in the equations. From (18), (21), (25), (26), (30) and (31) we see that the contour integrals which are the coefficients of μ and ϵ involve normal components at the ferrite periphery of vector functions of the field quantities, while the contour integral coefficient of κ involves tangential components of vector functions of field quantities.

Another point observed from (18) is that cavities in which the field potential (aside from the time dependence) is real (e.g. rectangular or TEM type cavities) are not suited for determination of κ since the basic κ integral, $\oint_{c'} \xi \nabla_t \xi^* \cdot d\mathbf{l}'$, vanishes. A polarization that is natural to the spin precession is needed and this is provided in the circular or square type cavities. In these cases the spatial field potential is complex and the κ integral does not vanish.

All results presented here apply also to the situation when the ferrite is lossy. In that case, μ, ϵ, κ and ω become complex in all formulas. A discussion of the Q of the cavity in this case is given in the various papers of Ref. 1.

APPENDIX I

Derivation of Equation (10)

Gauss' Theorem applied to the disc of Fig. 1 is

$$\int \text{div} \, (\xi \nabla_t \xi^*) \, d\tau' = \int (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0' \, dA'. \tag{45}$$

Since $\xi \nabla_t \xi^*$ is parallel to the surface S', there is no flux of this vector through S'. Thus, with $dA' = t \, dl'$ and $d\tau' = dS' \cdot t$ we have

$$\int \text{div} \, (\xi \nabla_t \xi^*) \, dS' = \oint_{c'} (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0' \, dl' = \int (\xi \nabla_t^2 \xi^* + |\nabla_t \xi|^2) \, dS'.$$

With $(\nabla_t^2 + k_c^2)\xi = 0$, this becomes

$$\int_{S'} |\nabla_t \xi|^2 \, dS' = k_c^2 \int_{S'} |\xi|^2 \, dS' + \oint_{C'} (\xi \nabla_t \xi^*) \cdot \mathbf{n}_0' \, dl', \qquad (46)$$

which is also applicable to C and S.

APPENDIX II

Derivation of Equation (11)

Consider two functions ξ_1 and ξ_0 which differ slightly, both satisfying the wave equation, and corresponding to the wave vectors $\mathbf{k}_{c_1}$ and $\mathbf{k}_c$, respectively:

$$(\nabla_t^2 + k_c^2)\xi_0 = 0,$$

$$(\nabla_t^2 + k_{c_1}^2)\xi_1 = 0,$$

$$\mathbf{k}_{c_1} = \mathbf{k}_c + d\mathbf{k}_c, \qquad (47)$$

$$k_{c_1}^2 = k_c^2 + 2\mathbf{k}_c \cdot d\mathbf{k}_c,$$

$$\xi_1 = \xi_0 + (\nabla_{k_c}\xi_0) \cdot d\mathbf{k}_c,$$

where ∇_{k_c} is the gradient operator in $\mathbf{k}_c$ space. We insert (47) into Green's identity which, for the geometry of Fig. 1, is

$$\int_{S'} (\xi_0^* \nabla_t^2 \xi_1 - \xi_1 \nabla_t^2 \xi_0^*) \, dS' = \oint_{C'} (\xi_0^* \nabla_t \xi_1 - \xi_1 \nabla_t \xi_0^*) \cdot \mathbf{n}_0' \, dl', \quad (48)$$

and equate coefficients of $d\mathbf{k}_c$ terms on both sides. Then in the limit as ξ_1 approaches $\xi_0 \equiv \xi$, and noting from (48) and the wave equation that

$$\oint_{C'} (\xi_0^* \nabla_t \xi_0 - \xi_0 \nabla_t \xi_0^*) \cdot \mathbf{n}_0' \, dl' = 0,$$

this procedure gives

$$2k_c^2 \int_{S'} |\xi|^2 \, dS' = \oint_{C'} [(\nabla_{k_c}\xi \cdot \mathbf{k}_c)(\nabla_t \xi^* \cdot \mathbf{n}_0') - \xi^* \nabla_t (\nabla_{k_c}\xi \cdot \mathbf{k}_c) \cdot \mathbf{n}_0'] \, dl', \quad (49)$$

which is again applicable to C and S. A one-dimensional form of this result is given by Sommerfeld.[4]

APPENDIX III

Derivation of Equation (12)

The following identity holds:

$$\nabla_t \times (\xi \nabla_t \xi^*) = \nabla_t \xi \times \nabla_t \xi^* + \xi \nabla_t \times \nabla_t \xi^* = \nabla_t \xi \times \nabla_t \xi^*, \quad (50)$$

so that

$$\int_{s'} (\nabla_t \xi \cdot \mathbf{z}_0 \times \nabla_t \xi^*) \, dS' = \int_{s'} \mathbf{z}_0 \cdot (\nabla_t \xi^* \times \nabla_t \xi) \, dS'$$

$$= \int_{s'} (\nabla_t \xi^* \times \nabla_t \xi) \cdot d\mathbf{S}' = \int_{s'} \nabla_t \times (\xi \nabla_t \xi^*) \cdot d\mathbf{S}' = \oint_{c'} (\xi \nabla_t \xi^*) d\mathbf{1} \cdot' \qquad (51)$$

where Stokes' theorem has been employed and $d\mathbf{S}' = \mathbf{z}_0 dS'$.

APPENDIX IV

Derivation of Magnetic Energy Stored in Ferrite, Equation (17)

We decompose the RF $\mathbf{H}_0$ field in the cavity (16) into a combination of two circular polarizations of opposite senses and possibly different amplitudes H_+ and H_-,

$$\mathbf{H}_0 = H_+ \boldsymbol{\delta}_+ + H_- \boldsymbol{\delta}_-, \qquad (52)$$

where $\boldsymbol{\delta}_+$ and $\boldsymbol{\delta}_-$ are two-dimensional column vectors describing circular polarizations of opposite senses and unit amplitudes:

$$\boldsymbol{\delta}_+ = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -j \end{pmatrix}, \boldsymbol{\delta}_- = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ j \end{pmatrix}. \qquad (53)$$

Both $\boldsymbol{\delta}_+$ and $\boldsymbol{\delta}_-$ satisfy the following relations:

$$\boldsymbol{\delta}_+{}^* = \boldsymbol{\delta}_-, \qquad \boldsymbol{\delta}_+ \cdot \boldsymbol{\delta}_+{}^* = 1, \qquad \boldsymbol{\delta}_- \cdot \boldsymbol{\delta}_-{}^* = 1, \qquad \boldsymbol{\delta}_+ \cdot \boldsymbol{\delta}_+ = 0,$$

$$\boldsymbol{\delta}_-{}^* = \boldsymbol{\delta}_+, \qquad \boldsymbol{\delta}_+ \cdot \boldsymbol{\delta}_- = 1, \qquad \boldsymbol{\delta}_- \cdot \boldsymbol{\delta}_+ = 1, \qquad \boldsymbol{\delta}_- \cdot \boldsymbol{\delta}_- = 0. \qquad (54)$$

Further from (15) we have

$$\chi \boldsymbol{\delta}_\pm = \chi_\pm \boldsymbol{\delta}_\pm,$$

$$\chi_\pm = (\mu - 1) \pm \kappa. \qquad (55)$$

Then it is easily shown that

$$\mathbf{M} \cdot \mathbf{H}_0{}^* = \chi_+ \mid H_+ \mid^2 + \chi_- \mid H_- \mid^2, \qquad (56)$$

where from (16) and (53)

$$H_{\pm} = \mathbf{H}_0 \cdot \mathbf{d}_{\pm}{}^*$$

$$= \frac{-j\beta \cos \beta z}{\sqrt{2}\omega\mu_0} \left[\left(\frac{\partial \psi}{\partial u} \pm j \frac{\partial \psi}{\partial v} \right) \text{ or } \frac{-k^2}{\beta^2} \left(-\frac{\partial \varphi}{\partial v} \pm j \frac{\partial \varphi}{\partial u} \right) \right] \tag{57}$$

for a ψ wave and φ wave, respectively. Then (56) and (57) give

$$\mathbf{M} \cdot \mathbf{H}_0{}^* = f \frac{\beta^2 \cos^2 \beta z}{\omega^2\mu_0{}^2} \left[(\mu - 1) \mid \nabla_t \xi \mid^2 + j\kappa \nabla_t \xi \cdot \mathbf{z}_0 \times \nabla_t \xi^* \right] \tag{58}$$

where $f = 1$ or $(k^2/\beta^2)^2$ for $\xi = \psi$ and $\xi = \varphi$, respectively. Equation (17) follows from an integration of (58) throughout the sample, with $d\tau' = dS' \, dz$.

REFERENCES

1. Artman, J. O. and Tannenwald, P. E., Measurement of Permeability Tensor in Ferrites, Phys. Rev., **91**, Aug. 15, 1953, p. 1014; Measurement of Susceptibility Tensor in Ferrites, J. Appl. Phys., **26**, Sept. 1955, p. 1124.
2. Rowen, J. H. and Von Aulock, W., Measurement of the Complex Tensor Permeability of Ferrites, Phys. Rev., **96**, Nov. 1954, p. 1151; Measurement of Dielectric Magnetic Properties of Ferromagnetic Material at Microwave Frequencies, B. S. T. J., **36**, March 1957, p. 427.
   Lax, B. and Berk, A. D., Convention Record of the Meeting of the I.R.E., March 1953 (unpublished).
   Beljers, H. G., Measurements on Gyromagnetic Resonance of a Ferrite Using Cavity Resonators, Physica, **14**, Feb. 1949, p. 629.
   Spencer, E. G., LeCraw, R. C. and Reggia, F., Circularly Polarized Cavities for Measurement of Tensor Permeabilities, J. Appl. Phys., **26**, March 1955, p. 354.
   Berk, A. D. and Lengyel, B. A., Magnetic Fields in Small Ferrite Bodies with Applications to Microwave Cavities Containing Such Bodies, Proc. I.R.E., **43**, Nov. 1955, p. 1587.
   Waldron, R. A., Proc. Inst. Elect. Eng., **104B**, No. 6, Oct.-Nov. 1956, pp. 307–315.
2. Casimir, H. B. G., On the Theory of Electromagnetic Waves in Resonant Cavities, Philips Res. Rept., **6**, 1951, pp. 162–182.
   Bethe, H. A. and Schwinger, J., N.D.R.C. Rept. D1-117, March 1943.
   Muller, J., Hochfr.-Tech. und Elektr., **54**, 1939, p. 157.
3. Jahnke, E. and Emde, F., *Tables of Functions*, 4th Ed., Dover Publishers, N. Y., 1945, p. 144.
4. Sommerfeld, A., *Partial Differential Equations*, Academic Press, N. Y., 1949, p. 310.

The Effects of Mode Filters on the Transmission Characteristics of Circular Electric Waves in a Circular Waveguide

By W. D. WARTERS

Departures from perfect geometry in a multi-mode circular waveguide used for circular electric wave transmission will cause coupling between the TE_{01} signal mode and the other propagating modes. Such coupling causes serious degradation of signal fidelity after long travel distances. These effects have been studied in a long 5-inch-diameter guide in the 9000-mc band, for effective pulse travel distances up to 12 miles. Mode filters have been developed which suppress all spuriously generated modes in this guide. It is found that the insertion of these mode filters at reasonable intervals along the waveguide reduces pulse distortion to a negligible level and smooths the variations in the loss versus frequency characteristic.

I. INTRODUCTION

A circular waveguide operated in the TE_{01} or circular electric mode is very attractive for use as a long-distance transmission medium.[1,2] The attenuation coefficient of the TE_{01} mode decreases monotonically with increasing frequency in a given waveguide so that, in theory, one can transmit information through such a system with as low a loss as desired merely by operating at a sufficiently high frequency. In practice, in order for one to take advantage of this low-loss property, the frequency must be so high that the waveguide is operating in the multimode region. This introduces the probability that coupling will occur between the low-loss TE_{01} signal mode and the other propagating modes if there are any deviations from perfect circularity and straightness in the guide. The coupling resulting from a series of imperfections, with the resultant transfer of power back and forth between TE_{01} and other modes, can cause serious degradation of the transmission properties of a long waveguide. These mode conversion and reconversion effects

657

present one of the most serious problems to the use of waveguide as a practical long-distance transmission medium.

Much research has already been done on this and other aspects of TE_{01} transmission properties. Discussions of earlier contributions are included in a paper by Miller[2] which views the over-all problems and possibilities of TE_{01} transmission systems and discusses in detail many specific problems, including mode conversion.

Several ways of combating mode conversion effects have been proposed. Good results can be obtained over narrow frequency bands by cancelling one imperfection against another by such actions as orienting individual pipe sections or adjusting the joints between sections. However, this approach destroys one of the most attractive features of waveguide as a transmission medium, its enormous bandwidth capabilities. An obvious solution is to eliminate the mode conversion by obtaining as nearly perfect a waveguide as possible. Unfortunately, the tolerances required to accomplish this in a practical waveguide appear to be unattainable. A more reasonable approach is to provide high loss in the waveguide to all modes except the desired TE_{01} mode, thereby attenuating the spuriously generated modes before reconversion to TE_{01} can occur. Ideally, this loss would be provided continuously along the waveguide, perhaps by using a modified guide structure such as a helix waveguide.[2,3] However, considerable improvement can be achieved in the performance of a regular solid copper guide by inserting discrete mode filtering structures at intervals along its length.

The purpose of the experiments described in this paper was (1) to investigate the seriousness of mode conversion effects in a long waveguide which had been erected as carefully as possible, (2) to develop a discrete filter which would be effective against the spurious modes generated from TE_{01} in this waveguide and (3) to study the TE_{01} transmission properties of the line with these filters inserted at various spacings.

II. OBSERVATIONS OF MODE CONVERSION EFFECTS

The waveguide used in these experiments was the 4.73-inch I.D., 5.00-inch O.D. installation at Bell Telephone Laboratories in Holmdel, N.J. It was constructed in 1946 for operation at 9000 mc, where it will support 40 modes. The theoretical TE_{01} attenuation constant at 9000 mc is about 2 db per mile. It is the same line used and described in earlier papers.[1,2]

The shuttle-pulse equipment developed by A. C. Beck[1] was used for the transmission tests. This system is shown in Fig. 1. A 0.1 microsecond pulse is introduced into the line and bounces back and forth between

shorting pistons at each end. The transmitter and receiver coupling holes are so small that the loss through them is negligible compared to the heat loss of the line. By changing the variable delay, which controls the scope sweep, one can observe the pulse on the scope at any desired time after it has entered the line and, therefore, after any desired number of trips up and down the line. The coupling holes respond to many of the 40 propagating modes, but, because of the high attenuation of all but TE_{01} , only that remains after several trips.

This system provides good conditions for studying mode conversion-reconversion phenomena in the line because the pulse traverses the same guide many times. Since identical conversions and reconversions occur during each trip, their effects become very pronounced after many trips. Also, if no loss to unwanted modes is provided at the ends of the line, an imperfection which causes a conversion from TE_{01} to some other mode X as the signal pulse travels down the line will cause a reconversion of the same magnitude from X back to TE_{01} when the spurious pulse X travels back up the line. This reconverted TE_{01} pulse will, in general, not be in phase with the main TE_{01} signal pulse which is also traveling back up the line. The phase difference depends on the identity of mode X, the operating frequency and the distance between the imperfection and the reflecting piston. By moving the piston, one can change this phase difference and observe the period of the interference with piston motion, thereby identifying mode X. Since the same process occurs each trip, the magnitude of the interference increases
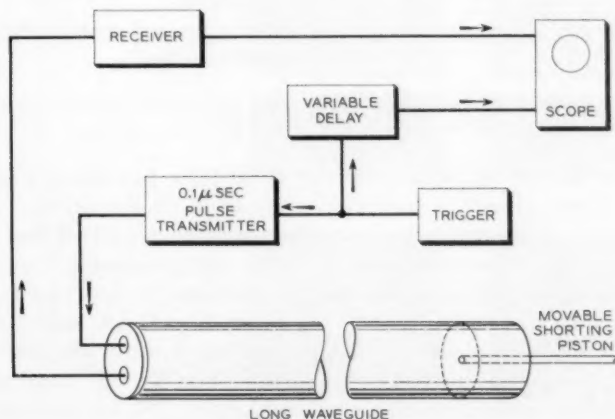

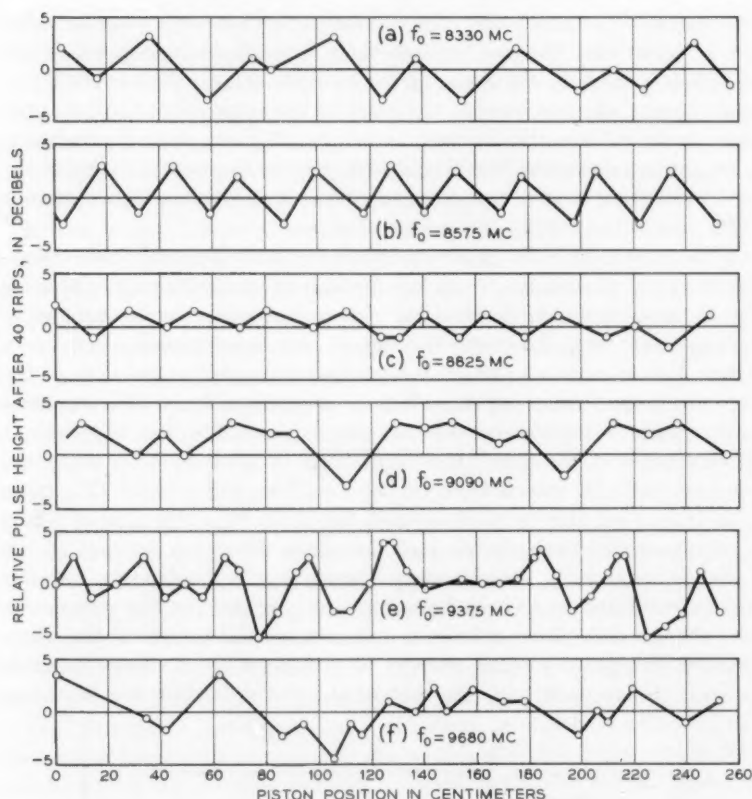Fig. 1 — Block diagram of pulse transmission test equipment.

Fig. 2 — Measured variation of TE_{01} pulse height with piston position after 40 round trips in a 425-foot waveguide with no filters.

with the number of round trips the signal pulse has traveled, so very small effects become visible after many trips. Of course, this process occurs for each imperfection in the line with the magnitude and piston period characteristic of that imperfection and the modes it generates. Thus, for a line with many random imperfections, we can expect a very complicated fluctuation of observed pulse height with piston motion when no mode filters are in the line. If effective filters are inserted in front of the piston, power which is converted on the way down the line will be absorbed and can no longer reconvert to TE_{01} on the way back. Therefore, the fluctuations with piston motion should disappear when mode filters are inserted. The same general argument explains the effects

observed when the operating frequency is changed and the piston is held stationary, except that, in this case, the electrical length of the waveguide is changed. Therefore, the effect of changing frequency is similar to that of linearly stretching or shrinking the guide at constant frequency.

The effect on the TE_{01} pulse of traveling many round trips in a waveguide with no filters is shown in Figs. 2 and 3. The guide was about 425 feet in length to the central piston position. Fig. 2 shows the measured variation of the pulse height with piston position after 40 trips at various frequencies. The available piston travel was about eight feet. Points were taken only at the extremes of the variations, so fine-grain accuracy was not obtained. Although some of the curves show regularities, the variation in general is more or less random, as is expected for a line with random imperfections. Photographs of the oscilloscope traces were taken at several frequencies at the piston positions giving the highest and the lowest observed pulse heights. Some of these photographs are shown in Fig. 3. Each trace shows three consecutive trips, the center trip having traveled the distance noted at the left of each pair of photographs. The receiver sensitivity has been adjusted in each case so that the pulses are the same size. Of course, the attenuation constant is actually much higher at the unfavorable piston positions, in some cases by a factor of 1.5.
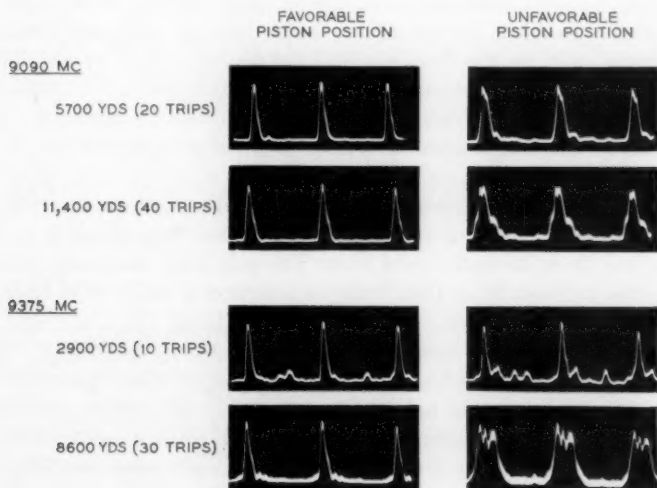

Fig. 3 — Oscillograms of pulses at good and bad piston positions for various travel distances in a 425-foot waveguide with no filters.

It is clear that the distortion of the pulse becomes very severe after many trips with the piston at the unfavorable position. This distortion occurs because the reconverted TE_{01} pulses, having traveled for a while in spurious modes, occur at different times and with different phases from those of the signal pulse. The small pulses along the baseline in the 9375-mc pictures at 10 trips are mostly modes other than TE_{01} which were introduced by the transmitting coupler and have not yet died out. However, all distortions of the pulses at larger distances are caused by mode conversion-reconversion interference.

Obviously, mode conversion has important effects on the transmission characteristics of this line. The next step in the investigation was to identify the spurious modes present and to discover a mode-filtering structure which would provide high loss to them.

III. DEVELOPMENT OF AN EFFECTIVE MODE FILTER FOR THE 5-INCH LINE

Earlier work by A. P. King and others at Bell Laboratories has resulted in several mode filter designs.[2] Some of these are illustrated in Fig. 4. The radial card filter consists of resistive sheets mounted in foamed polystyrene along diametral planes of the guide. The TE_{0n} E-lines, being concentric circles, are always normal to the sheets. All other modes have either a radial or longitudinal component (or both) of electric field and therefore suffer loss in such a device. The spaced-ring line and helical line are structures which provide good guide-wall conductivity only in the circumferential direction, which is all that the circular electric wave requires, and poor conductivity in the longitudinal direction. Since all modes except the circular electric modes have wall current in the longitudinal direction, they suffer increased loss in such a structure.

It was felt that the spaced-ring and helical structures would be best suited to the purposes of these experiments, since they should have lower TE_{01} loss than the radial card filter. The polyfoam mounting raises the TE_{01} loss sufficiently so that frequent insertion of radial card filters in a long line is not practical.

Since the important spurious modes in the 5-inch line were not as yet identified, it was decided to perform preliminary studies on the effectiveness of spaced-ring and helical filters, using modes which were readily generated with available apparatus. Therefore, an experiment was performed to measure the TE_{11} and TM_{11} losses and the TE_{11}-TM_{11} conversions in these structures. The TM_{11} mode is known to be a serious problem in TE_{01} transmission around bends.[2, 4]

The experimental setup is sketched in Fig. 5. The pulse transmitter delivers a short pulse 0.006 microseconds in length at a 100-kc rate. Both this transmitter and the associated receiving equipment were developed and built by A. C. Beck and G. D. Mandeville of the Laboratories.[5] The filter under test is mounted at one end of about 300 feet of the 5-inch line, and the movable shorting piston is located inside the filter. Test pulses are then transmitted and received from the other end of the line by means of pure mode transducers.[6] With the setup as shown, the transmitted and received pulses can be in either the same or different modes. The short pulse gives enough time resolution for the modes of interest to be separated completely after 300 feet of travel

RESISTIVE SHEETS IN
POLYFOAM MOUNTING

(a) RADIAL CARD FILTER

SPACED COPPER RINGS

(b) SPACED RING LINE

LOSSY JACKET

FINE WIRE WOUND
IN CLOSELY-
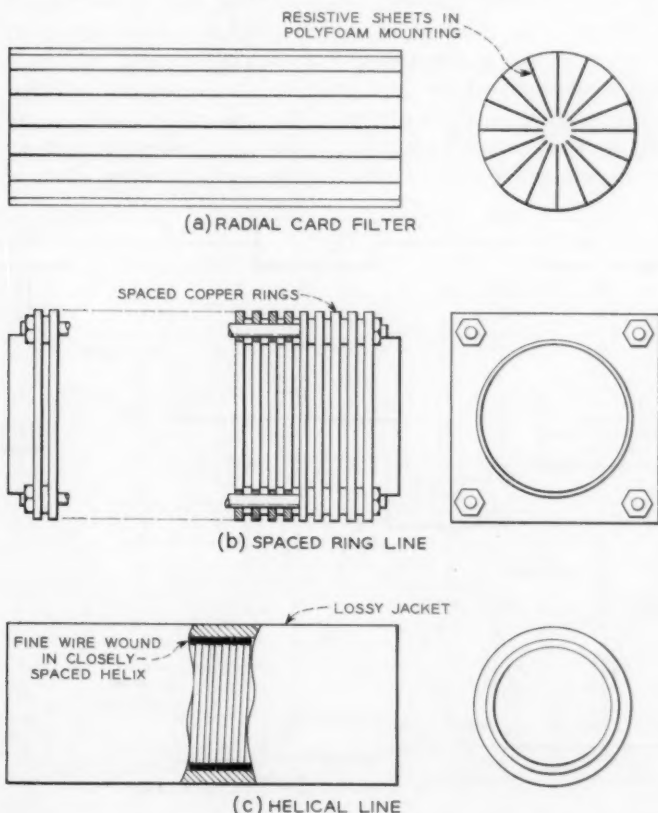SPACED HELIX

(c) HELICAL LINE

Fig. 4 — Several kinds of discrete mode filters.

because of their different group velocities. Changing the position of the piston changes the exposed length of the filter under test. Because of the reflection in the piston, the effect of a filter of twice the length of the exposed section is measured. Normalization of the data to zero filter length is easily done by placing the piston in front of the filter.

The results of measurements of TM_{11} and TE_{11} losses and of TE_{11}-TM_{11} conversion versus filter length in a spaced-ring line are plotted in Fig. 6. The filter used consisted of $\frac{1}{16}$-inch-thick copper rings separated by $\frac{1}{16}$-inch air spaces. It can be seen that the interactions are fairly complicated. Recent work by Morgan and Young on the theory of helix waveguide[3] indicates that a set of noncoupled propagating modes can exist in a spaced-ring or helical structure, but that their field configurations are quite different from those of the modes found in a solid circular guide. Therefore, one expects the insertion of such a filter in a solid guide to produce conversions at the first guide-to-filter boundary between the incident solid-guide mode to several of the helix-guide modes, as well as conversions at the second boundary from these helix-guide modes back to several solid-guide modes, including the incident mode. Each of the
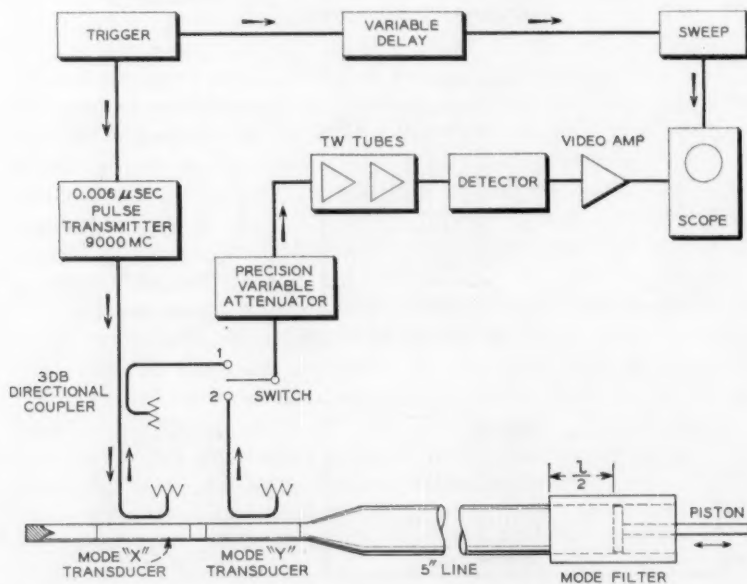

Fig. 5 — Block diagram of equipment for mode losses and conversions in mode filters, using the 9000-mc short pulser.

solid-guide modes departing from the filter will consist of several components, each produced from a different one of the modes inside the filter and with a relative phase that can be changed by changing the length of the filter. Therefore, an involved picture of loss versus filter length for the various modes is to be expected. A much simplified analysis of this interaction, using only two modes in the solid guide and two in the filter and disregarding reflections, is outlined in the Appendix. This solution gives fairly good agreement with the observed curves, and has been used to obtain the approximate attenuation constants noted on Figs. 6 and 7.

Spaced-ring lines of various spacings and ring thicknesses and with both air- and neoprene-filled spaces were tested in similar fashion. Helical lines wound of several wire sizes and with various jacket materials were also built and measured. Fig. 7 shows the results obtained with a helix wound of No. 30 Formvar-coated wire and coated with a carbon-loaded casting cement. This helix gave the best performance of the several filter designs tested. Therefore, several copies of it were made for use in the tests described below.

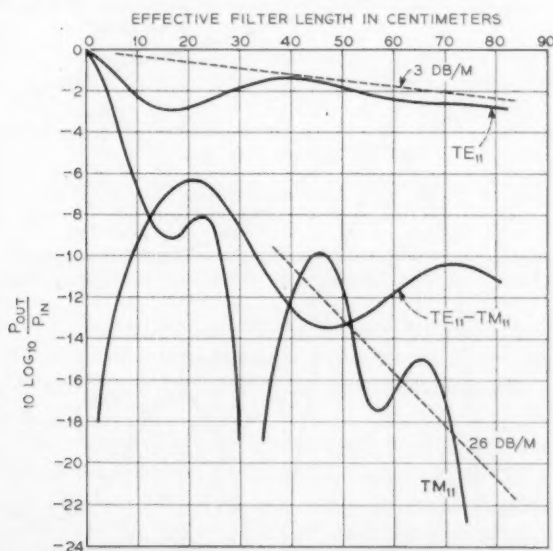The effects on the TE_{01} pulse of the 425-foot line with several 18-

Fig. 6 — Measured TE_{11} loss, TM_{11} loss and TE_{11}-TM_{11} conversion vs. filter length in an air-spaced, spaced-ring mode filter.

inch helical filters spaced uniformly along its length are shown in Figs. 8 and 9. These measurements were taken in the identical fashion as the no-filter results given in Figs. 2 and 3.

Fig. 8 shows the measured pulse height variation versus piston position after 40 trips, with nine helices in the line. These curves were taken at the same frequencies and are plotted on the same scale as the curves of Fig. 2. The filters produce a marked reduction in the pulse height fluctuations at 8330, 8825 and 9090 mc. At 8575, 9375 and 9680 mc, the fluctuations are only slightly lessened and have become almost periodic. This regularity indicates mode conversion-reconversion interference between TE_{01} and a single other mode. The period of the variation at all these frequencies is exactly that to be expected from a TE_{01}-TE_{12} conversion-reconversion process.

Fig. 9 includes photographs of the pulses at the same frequencies and piston positions as Fig. 3, after eight helices were inserted in the line. The distortion is greatly reduced, and at 9090 mc, where the pulse height
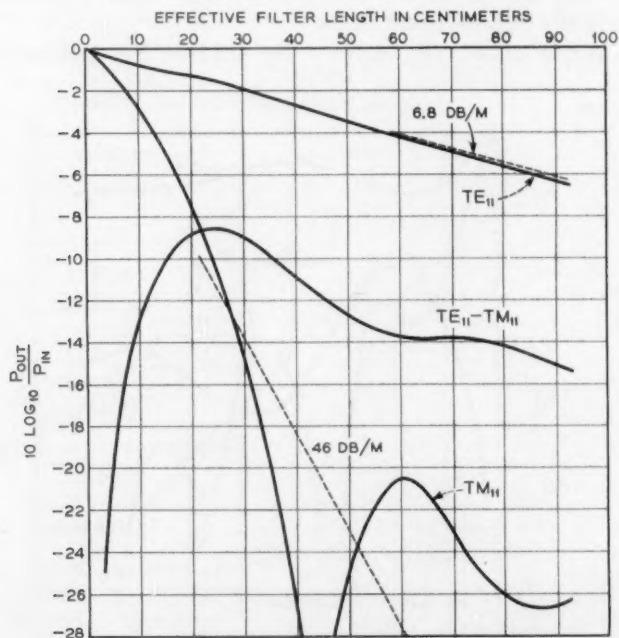

Fig. 7 — Measured TE_{11} loss, TM_{11} loss and TE_{11}-TM_{11} conversion vs. filter length in a helix mode filter constructed of No. 30 wire with a lossy jacket.
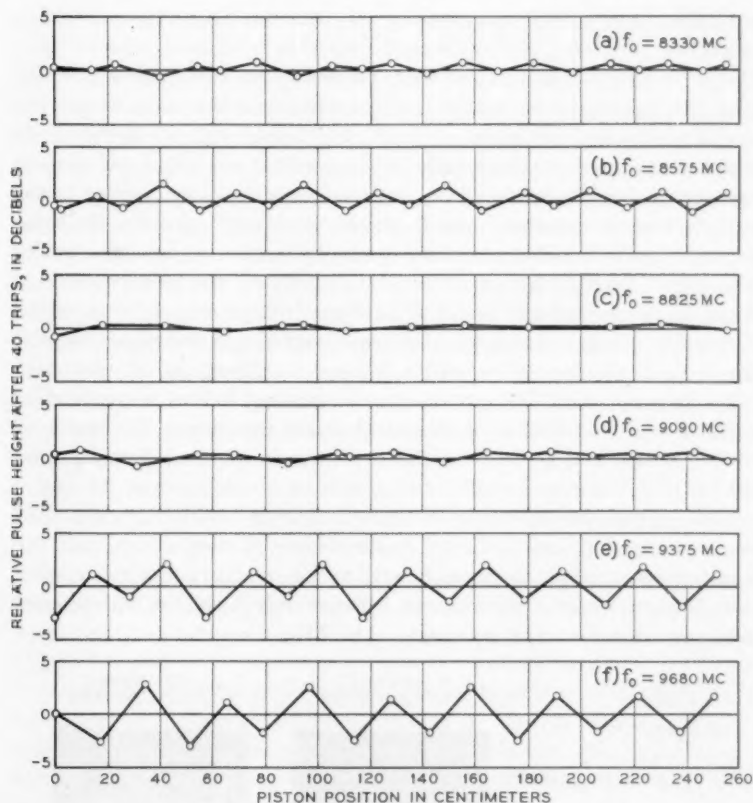
Fig. 8 — Measured variation of TE_{01} pulse height with piston position after 40 round trips in a 425-foot waveguide with nine helix mode filters.

fluctuation was greatly reduced, the difference between favorable and unfavorable piston settings is practically eliminated. Note that, because of this improvement, we were able to observe the pulses after more trips than in Fig. 3.

We conclude that the helical filters used in this experiment were effective against all spuriously generated modes except TE_{12}. At frequencies where the net TE_{01}-TE_{12} conversion in the waveguide is small, presumably because the electrical length between conversion centers is unfavorable to TE_{12} build-up, the pulse height and shape vary little with piston setting after the addition of the helices to the line. At frequencies where the TE_{01}-TE_{12} conversion is strong, the pulse shape

and height vary with piston position and with the TE_{01}-TE_{12} conversion period.

The mode TE_{12} arises from such defects in the waveguide as offset axes, tilted joints or curvature in the longitudinal direction. It has the second lowest loss of all the modes in the 5-inch pipe at 9000 mc, its attenuation constant being only $2\frac{1}{2}$ times that of TE_{01}. Its field is concentrated in the center of the guide and is but loosely coupled to the wall, producing relatively small longitudinal wall currents. For this reason, it is not affected as greatly by the helical line as are other modes with higher longitudinal currents. The TE_{12} loss in the helices described above was measured and found to be about 1 db per meter, or about 0.5 db per filter. The recent work of Morgan and Young indicates that this loss could have been increased by the use of different jacket materials.

To improve their TE_{12} performance, existing helical sections were modified by the addition of diametral resistance sheets. The resulting filter is shown in Fig. 10. The sheets are made by coating thin Mylar film on both sides with resistive material, then cutting from the film a section shaped like two isoceles triangles joined at their apexes. This section is twisted and mounted inside a helix in such a way that the plane of one triangle is perpendicular to the plane of the other, each plane being defined by the axis and a diameter of the guide. This twisted sheet provides loss to all polarizations of TE_{12}.
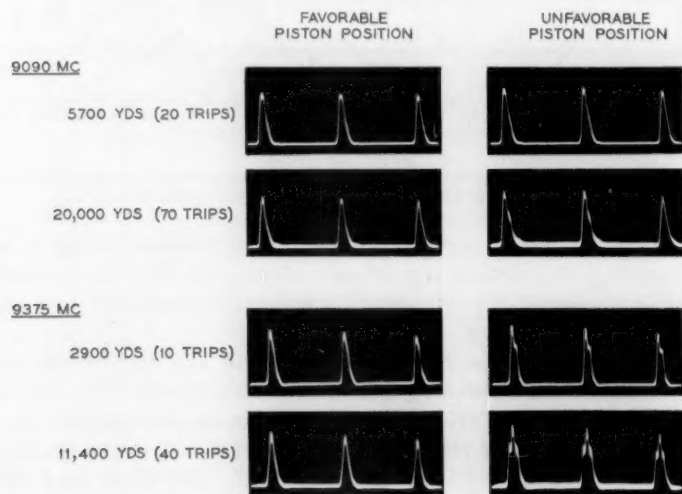

Fig. 9 — Oscillograms of pulses at good and bad piston positions for various travel distances in a 425-foot waveguide with eight helix mode filters.

Since it was impossible to run the piston inside the resistance-sheet helical filters, a slightly different measuring technique from that used with the spaced-ring and helical lines was employed to determine the optimum surface resistivity of the sheet. Instead of the test section being at the far end of the line (see Fig. 5), it was placed at the near end, just beyond the transducers. Using this system with the short-pulse equipment, modes produced at the filter by conversion from the incident mode are resolved in time by the trip down the line and back. On the return trip through the filter, each of these modes will again suffer various conversions, so that there will be pulses in several different modes at each of several times, each time corresponding to the travel time of a particular mode produced at the first trip through the filter. Therefore, by sending and receiving through pure-mode transducers and measuring travel times, one knows the transmitted mode, the mode in which the power traveled and the received mode — and quite an assortment of mode loss and conversion data can be obtained.

Fig. 11 shows the measured TE_{11} and TE_{12} losses and TE_{11}-TM_{11}, TE_{11}-TE_{12} and TE_{11}-TE_{31} conversions in a resistive-sheet helix versus the RF surface resistivity of the film. The TE_{11} and TE_{12} losses show definite maxima with resistivity. From these data it was decided that a film resistance of 200 to 300 ohms per square gave optimum performance. Accordingly, films were made in this range and mounted in the available helices.

A comparison of the effectiveness of the helical filters with that of the resistive-sheet helices at 9375 mc is shown in Fig. 12. With two regular helices in front of the movable piston, the 40th-trip TE_{01} pulse height varies about ±3 db with piston position, exhibiting the TE_{01}-TE_{12} period. With one resistive-sheet helix before the piston, the variations are reduced to ±0.5 db, which is within experimental error.
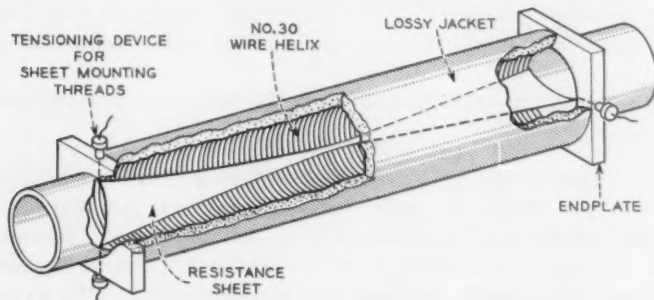

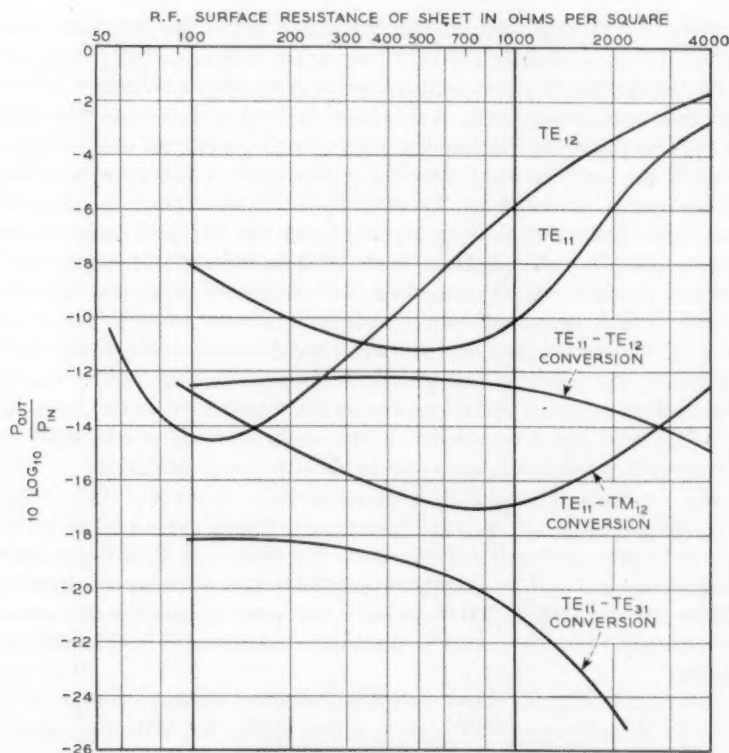Fig. 10 — A resistance-sheet helix mode filter.

Fig. 11 — Measured losses and conversions vs. sheet surface resistance for several modes in resistance-sheet helix mode filters.

We conclude that the resistive-sheet helices are effective in suppressing all modes which are spuriously generated from TE_{01} in the 5-inch wave-guide. They were used in the remainder of the work reported below.

IV. MEASUREMENTS OF ATTENUATION VERSUS FREQUENCY WITH VARIOUS FILTER SPACINGS

In a practical use of waveguide as a long-distance transmission medium, the effects of mode conversion-reconversion will become apparent not by moving pistons, but by variations in the attenuation versus frequency characteristic of the waveguide. For a broadband system, it is obviously desirable that this characteristic be flat over wide regions of the spectrum to be used. If mode conversion is present,

one expects a more-or-less chaotic variation of the attenuation versus frequency, depending on the relative positions and phases of the converting centers, as described earlier.

The insertion of discrete mode filters at regular intervals in a long waveguide divides the line into a series of short independent sections, since the conversion occurring in one section cannot cause reconversion in another. At a given frequency, the attenuation constant of the line is the average of the attenuation constants of these independent sections. This average will vary about some mean value as the frequency is changed. By decreasing the filter spacing, and thereby increasing the number of independent sections in a given line, the average is taken over a larger sample and its variance about its mean should decrease. Also, since the maximum distance over which interactions can occur will be reduced, the average attenuation constant will vary less rapidly with frequency. Therefore, we expect that the attenuation versus frequency characteristic of a given line will show smaller and less rapid variations about a constant mean value as the filter spacing is decreased.

Measurements of this kind have been made in a 500-foot length of the 5-inch line. The results are shown in Fig. 13. Points were taken every 25 mc in frequency, except in the vicinity of 8800 mc, where a resonance in the sending apparatus at the TM_{42} mode cutoff frequency interfered with the measurement. The 0.1-microsecond pulse which was used has a width of about 20 mc, and the reproducibility of the pulse center frequency is ± 5 or 10 mc; so the measurements are about as fine-grained as is possible with this equipment. The TE_{01} loss of the resistance-sheet
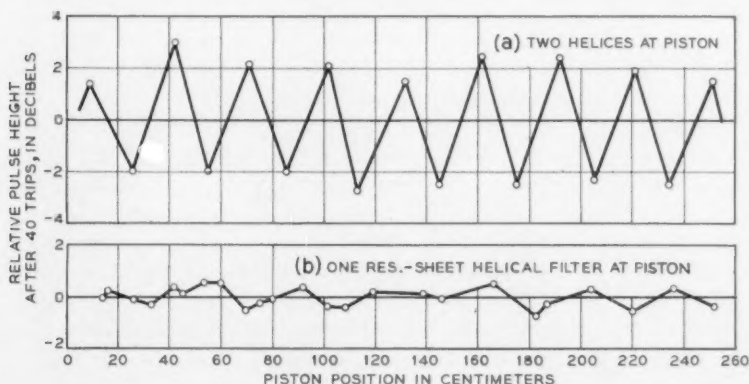

Fig. 12 — Comparison of effectiveness of helix and resistance-sheet helix mode filters at 9375 mc. Measured pulse height vs. piston position variations after transmission through two helices or one resistance-sheet helix.
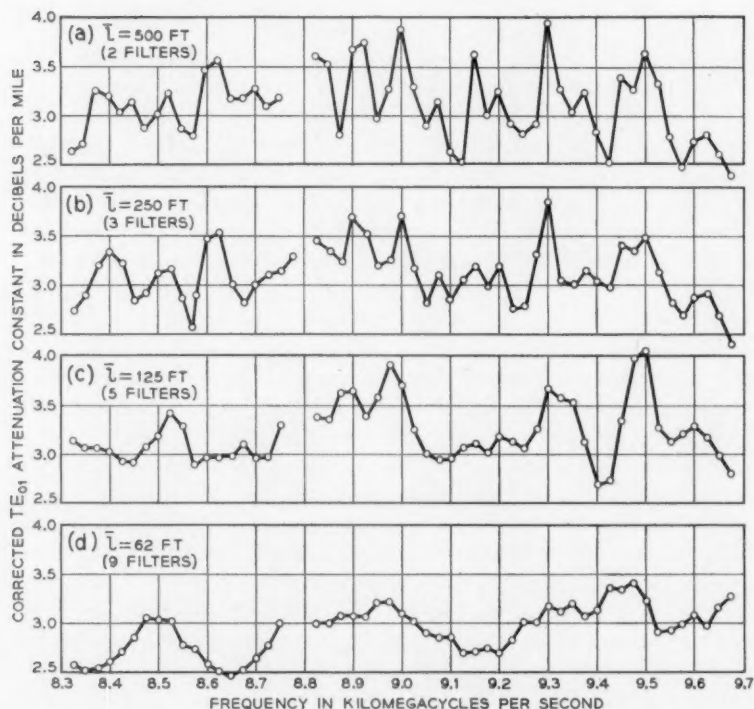
Fig. 13 — Measured TE_{01} attenuation constant vs. frequency in a 500-foot waveguide with resistance-sheet helix filters at various spacings.

helices, 0.027 db per filter, has been subtracted from these results to allow a direct comparison of the mode conversion effects.

Curve (a) of Fig. 13 shows the measured TE_{01} attenuation constant versus frequency with a resistance-sheet helical filter at each end of the 500-foot line. This arrangement decouples the mode conversion effects of one trip from the next trip and gives the true attenuation constant of the line as a whole. Curve (b) shows the measured TE_{01} attenuation constant versus frequency with a third mode filter inserted at the middle of the line, giving an average filter spacing of 250 feet. Curve (c) gives the results with an average spacing of 125 feet, and (d) gives those with an average spacing of 62 feet. As the filter spacing is decreased, the fluctuations in the attenuation constant versus frequency decrease in magnitude, as expected, and the peaks broaden out so that the curves become much smoother.

For curve (d), with an average filter spacing of 62 feet, the 500-foot line was divided into eight supposedly independent sections by the nine mode filters used. To determine whether these eight sections indeed acted independently, a further experiment was performed. The results are shown in Fig. 14. The line was divided in the middle and the two halves were measured separately. These measurements are shown as
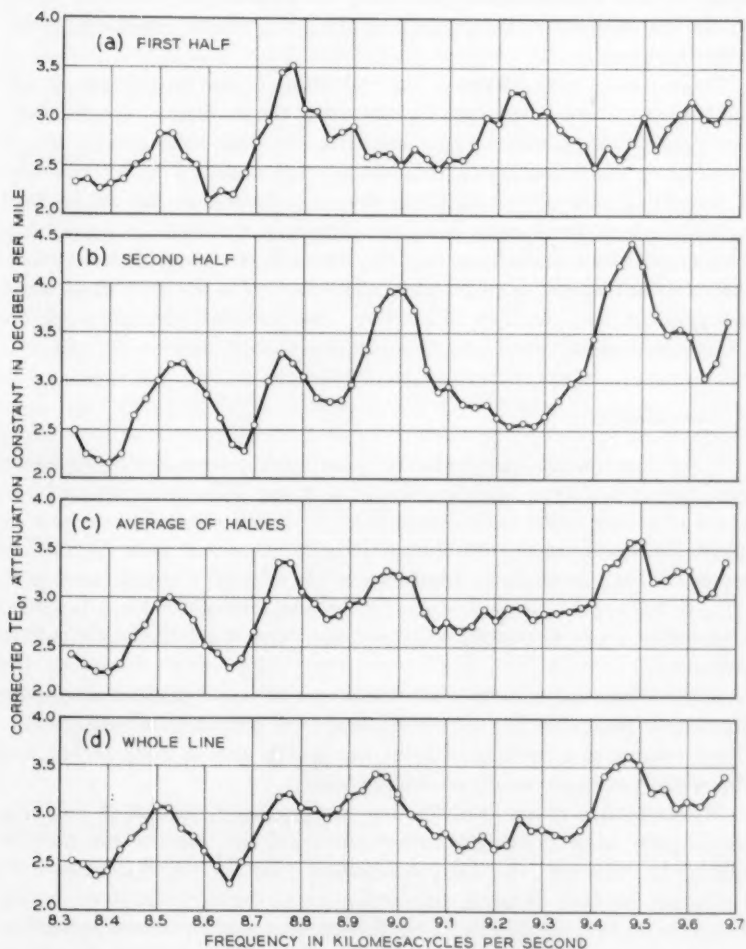

Fig. 14 — Measured TE_{01} attenuation constant vs. frequency for parts and for whole of 500-foot waveguide.

curves (a) and (b); they were made with a modified transmitting and receiving coupler so that the TM_{42} cutoff resonance at 8800 mc no longer appears. Although these curves are quite dissimilar, their average at each experimental frequency, which is plotted in curve (c), is in excellent agreement with the results obtained for the whole line, shown in curve (d). This agreement indicates that the individual line sections between filters do act independently and that the attenuation constant of the whole line is indeed the average of the attenuation constants of the individual sections.

These curves also illustrate the reduction in the magnitude of the fluctuations which is obtained by increasing the number of line sections. One expects that a much longer line with the same filter spacing would show much less variation of TE_{01} attenuation constant with frequency. It should be pointed out that this does not mean that the transmitted signal through the longer line will show less variation in amplitude. Elementary statistics tells us that the standard deviation of the attenuation constant should decrease as $1/\sqrt{n}$, where n is the number of equal independent line sections. Therefore, the standard deviation of the transmitted power level (expressed in db) should increase as $\sqrt{n}$.

V. CONCLUSION

These experiments have indicated that mode conversion-reconversion phenomena can have considerable effect on the TE_{01} transmission properties of a long multi-mode waveguide. It is very desirable to provide additional loss to modes other than TE_{01} by the use of mode filters. The mode TE_{12} is a particularly troublesome one because it is generated from TE_{01} in the types of imperfections likely to be encountered in a practical waveguide, has a low attenuation constant and is relatively difficult to suppress.

Mode filters have been built which are effective against all modes spuriously generated in a 5-inch waveguide at 9000 mc, including TE_{12}. They consist of a section of helix waveguide with a lossy jacket and diametral resistance sheets mounted inside.

With effective filters in a 500-foot waveguide, distortion of the TE_{01} signal pulse after it has traveled distances of the order of ten miles is negligible. However, the TE_{01} attenuation constant still fluctuates with frequency because of mode conversion-reconversion interactions in the line sections between filters. These fluctuations may be reduced considerably by decreasing the spacing between filters in the line. A spacing of

about 60 feet was found to smooth the attenuation versus frequency curve of a 500-foot waveguide sufficiently for it to be reasonably flat over bandwidths of the order of 100 mc.

ACKNOWLEDGMENTS

The continued assistance of G. D. Mandeville throughout these experiments is gratefully acknowledged. The author also wishes to thank his many co-workers at the Holmdel Laboratory who contributed helpful suggestions and discussions.

APPENDIX

We shall discuss here the junction between a solid waveguide and a helix waveguide under the simplifying assumptions that (1) the solution can be written in terms of only two solid-guide modes and two helix-guide modes; (2) there are no reflections from the junction and (3) the junction is lossless and reciprocal.

Under these assumptions one can write a 2×2 transfer matrix $[T]$ to represent the transmission from solid to helix guide. Transmission from helix to solid guide is then given by $[\tilde{T}]$, the transpose of $[T]$. Therefore the transmission through a helix-guide section of length l inserted in a solid guide is given by the matrix equations

$$[O] = [S(l)][E]$$

$$[S(l)] = [\tilde{T}] \begin{bmatrix} e^{-\Gamma_x l} & 0 \\ 0 & e^{-\Gamma_y l} \end{bmatrix} [T], \tag{1}$$

where $[O]$ and $[E]$ are column matrices representing the normalized output and input mode amplitudes, respectively, and Γ_x and Γ_y are the propagation constants in the helix guide of the two helix-guide modes.

When $l = 0$, we expect $[O] = [E]$, so that

$$[\tilde{T}][T] = [1]. \tag{2}$$

Conservation of energy requires that

$$[\tilde{T}][T^*] = [1], \tag{3}$$

where the asterisk denotes complex conjugate.

Equations (2) and (3) enable one to write the components of $[T]$ in

the form

$$[T] = \begin{bmatrix} \sqrt{1 - b^2} & b \\ -b & \sqrt{1 - b^2} \end{bmatrix},$$

so that

$$[S(l)] = \begin{bmatrix} (1 - b^2)e^{-\Gamma_x l} + b^2 e^{-\Gamma_y l} & b\sqrt{1 - b^2}(e^{-\Gamma_x l} - e^{-\Gamma_y l}) \\ b\sqrt{1 - b^2}(e^{-\Gamma_x l} - e^{-\Gamma_y l}) & b^2 e^{-\Gamma_x l} + (1 - b^2)e^{-\Gamma_y l} \end{bmatrix}.$$

The absolute squares of the components of $[S(l)]$ represent the power transmission and conversions for the solid-guide modes resulting from the presence of the helix section. After some algebraic manipulation, they may be written

$$| S_{11}(l) |^2 = 2b^2(1 - b^2)e^{-(\alpha_x + \alpha_y)l} [\cosh (\Delta\alpha l - \varphi) + \cos \Delta\beta l]$$
$$| S_{22}(l) |^2 = 2b^2(1 - b^2)e^{-(\alpha_x + \alpha_y)l} [\cosh (\Delta\alpha l + \varphi) + \cos \Delta\beta l] \qquad (4)$$
$$| S_{12}(l) |^2 = 2b^2(1 - b^2)e^{-(\alpha_x + \alpha_y)l} [\cosh \Delta\alpha l - \cos \Delta\beta l]$$

where $\Gamma_x = \alpha_x + j\beta_x$, $\Gamma_y = \alpha_y + j\beta_y$,
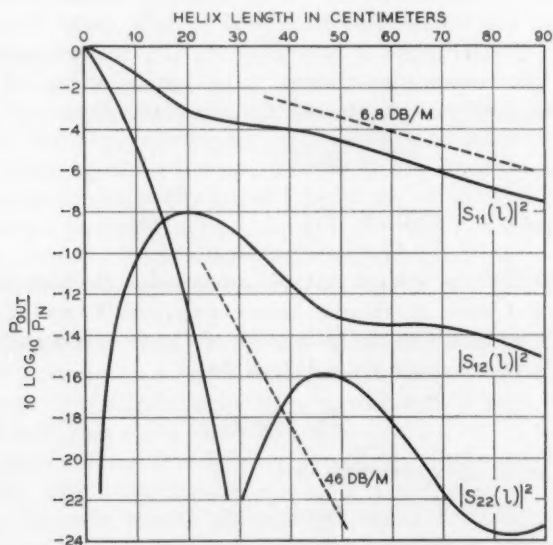

Fig. 15 — Plots of Equations (4) for $b^2 = 0.852$, $\alpha_x = 6.8$ db per meter, $\alpha_y = 46$ db per meter and $\Delta\beta = 10.6$ radians per meter.

and

$$\Delta\alpha = \alpha_y - \alpha_x$$

$$\Delta\beta = \beta_y - \beta_x$$

$$\varphi = \log_e \frac{b^2}{1 - b^2}.$$

Equations (4) are plotted in Fig. 15 for the values of the parameters which give the best fit to the TE_{11}-TM_{11} results for the helix of Fig. 7. A comparison of Figs. 7 and 15 shows good qualitative agreement. However, the lack of good quantitative agreement would indicate that the interaction is more complex than the simple case of the above analysis.

REFERENCES

1. Miller, S. E. and Beck, A. C., Low-Loss Waveguide Transmission, Proc. I.R.E., **41,** March, 1953, pp. 348–358.
2. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., **33,** November, 1954, pp. 1209–1266.
3. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** November, 1956, pp. 1347–1384.
4. Jouget, M., Effects of Curvature on the Propagation of Electromagnetic Waves in Guides of Circular Cross Section, Cables and Transmission (Paris), **1,** No. 2, July, 1947, pp. 133–153.
5. Beck, A. C. and Mandeville, G. D., Microwave Traveling-Wave Tube Milli-microsecond Pulse Generators, I.R.E. Transactions on Microwave Theory and Techniques, **MTT-3,** December, 1955, pp. 48–51.
6. Miller, S. E., Coupled-Wave Theory and Waveguide Applications, B.S.T.J., **33,** May, 1954, pp. 661–720.

Research Models of Helix Waveguides

By C. F. P. ROSE

A satisfactory technique of constructing research models of helix wave-guides has been evolved. A helix of insulated copper wire is wound on a mandrel and surrounded by a sheath of lossy material which, in turn, is covered by an outer plastic jacket reinforced with fiberglass. Units with 2-inch inside diameters have been made in lengths of 18 and 112 inches with observed circular-electric-wave losses 90 per cent above theory for an ideal copper tube at 55,000 mc.

In short lengths, the helix waveguide provides a useful component known as a mode filter which may be inserted at intervals in a solid-wall waveguide transmission line. The long lengths are of interest in producing an all-helix waveguide, which is expected to have superior transmission properties.

I. INTRODUCTION

A cylindrical metallic waveguide provides a unique transmission line for a circular electric wave having a wavelength less than one third the inside diameter of the waveguide. Under such conditions the lowest circular electric wave, TE_{01} , has less loss than any other mode. Furthermore, for perfectly round straight pipe, the losses for this mode decrease indefinitely as the frequency increases. Unlike all other modes, the circular electric modes induce no longitudinal wall currents. The currents which do flow are in the circumferential direction and result from the action of the pipe in restricting and directing the propagating energy.

Rigorous theory and experimental data substantiating the validity of the above statements have been presented elsewhere. (Refs. 1 through 6). The usefulness of these properties has provided the stimulus for a comprehensive study by Bell Telephone Laboratories at Holmdel, N. J. For this purpose, a round waveguide having a 2-inch inside diameter has been selected, and it is anticipated that its transmission will be useful over the band from 35,000 to 75,000 megacycles. Aside from economic considerations, the penalty for using a larger diameter is that the number of parasitic modes which may propagate is greatly increased and the physical tolerances become more critical.

679

There are numerous ways by which parasitic modes may be generated. Some of the most troublesome causes are axial discontinuities such as sharp bends, steps at junctions and ovality inside the pipe. Even though such mechanical irregularities are reduced to a practical minimum, it has been shown that the residue still produces electrical effects which are more than can be tolerated, (Ref. 5, pp. 1227–1229). One method of reducing effects of unwanted modes is to add mode filters at intervals along the line. An early version of such a device (Ref. 6, p. 1127), shown in Fig. 1, consisted of a succession of transverse copper rings separated from each other by spacer rings of lossy dielectric. This provides poor longitudinal conductivity but good circumferential conductivity, so that the losses for the circular electric wave group remain low while those for other modes become substantial. Since the unwanted energy is dissipated in the lossy medium, there is no possibility of reconversion to the circular electric mode. In the absence of such mode filters, the intermode effects cause serious distortion of the signal wave, because of a difference between the phase constants of the interfering mode and the signal mode, (Ref. 5, pp. 1230–1239).

A structure analogous to the one briefly described above but easier to construct has been developed and is illustrated in the cutaway view of Fig. 2. The completed structure is best described as a sheathed helix waveguide. The inner waveguide surface is a closely wound helix of insulated copper wire which is substituted for the spaced rings of its predecessor. The helix is surrounded by a lossy sheath and this, in turn,
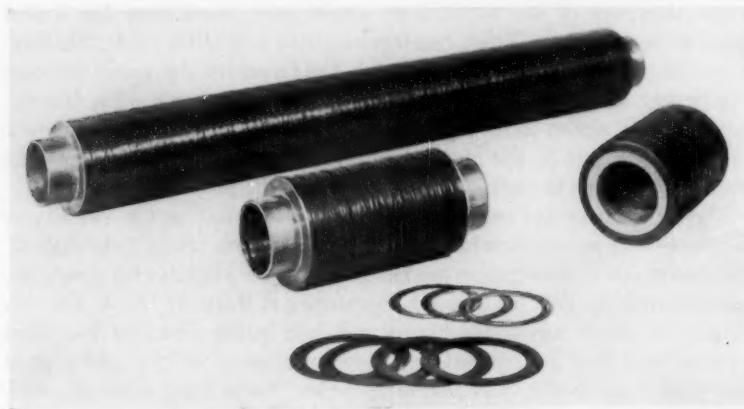

Fig. 1 — Spaced-disk mode filters.

Fig. 2 — 18-inch helix waveguide or mode filter.

is encased in a fiberglass-reinforced plastic jacket. Eighteen-inch lengths of such a structure offer a useful waveguide component for the laboratory and a device for reducing the effects of unwanted modes resulting from imperfections in an otherwise solid-wall waveguide. We are therefore interested in these short lengths for use as mode filters and in longer lengths for assembly into an all-helix waveguide system. The longer units were limited to 112 inches, since this is the present maximum convenient length from a constructional viewpoint. The following paragraphs tell how such units have been constructed.

II. CONSTRUCTION OF HELIX WAVEGUIDE MODE FILTERS

As mentioned previously, the mechanical irregularities in solid-wall waveguide must be minimized. For instance, for operation at 55,000 megacycles it appears desirable and practical to call for an over-all tolerance, including ovality, of ±0.001 inch for a nominal inside pipe diameter of 2.000 inches. At joints, a misalignment producing a step of greater than 0.0005 inch is undesirable. Even with these rigid degrees of tolerance in solid-wall waveguides, the introduction of parasitic modes is such that mode filtering is an additional requisite. The tolerances for the mode filters are just as exacting. Therefore, the casting of these glass-reinforced plastic units must start with a precise mold or mandrel. We have used one made of stainless steel tubing, centerless ground and polished to an outside diameter of 1.998 inches ±0.0005 inch. This nominal diameter for the mandrel allows 0.002 inch curing growth before the inside diameter of the finished product reaches its desired size. Mandrels having a highly polished finish of chrome or stainless steel are preferred, since they resist corrosion by atmospheric elements.
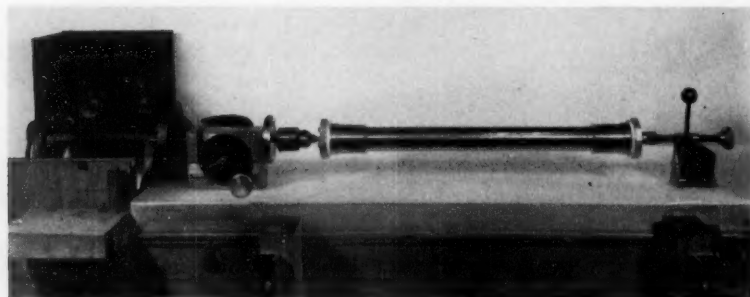
Fig. 3 — Mandrel and machine for winding mode filters.

The mandrel is mounted in a winding machine as shown in Fig. 3. This machine rotates the work while the helix and successive laminated jackets are hand-layered in a manner described subsequently and illustrated in Fig. 4.

After a suitable mold release agent has been applied to the mandrel, the helix is closely wound with plastic-insulated No. 37 copper wire of 0.005 inch over-all diameter. A uniform and appropriate spacing is provided between copper surfaces by the insulation on the wire. Two copper rings terminate the ends of the helix and provide abutment surfaces for joining to solid copper pipe.

After the helix is thoroughly degreased, it is covered with a fiberglass-reinforced plastic jacket consisting of three turns of 0.0015-inch thick woven fiberglass cloth which is impregnated with a thermosetting epoxy resin during lamination. The curing agent used for the epoxy resin is toxic and requires special handling to minimize dermatitis, but its thermosetting characteristics are preferred to the less toxic thermoplastic polyester resins because closer dimensional tolerances can be obtained. The developed wall thickness of this inner jacket is approximately 0.005 inch. The jacket performs two functions. It affords a good mechanical bond to the helix and it provides a transformation of the surface impedance of the lossy jacket presented to the longitudinal currents through the helix.

The lossy jacket is laminated next. It consists of a few turns of tin-oxide coated fiberglass cloth wrapped over the inner jacket and laminated with epoxy resin. This sheath around the helix is the medium in which the energy of unwanted modes is dissipated.

The intermediate jacket surrounding the lossy jacket is laminated with sufficient turns of clear glass cloth to develop a diameter equal to the outer diameter of the copper rings.

The helix waveguide is completed with an outer fiberglass-reinforced plastic jacket which has an over-all diameter of 3 inches. Its purpose is to provide strength. This $\frac{7}{16}$-inch wall jacket consists of additional turns of 0.005-inch clear fiberglass woven cloth which is hand-layered and laminated with epoxy resin extending the entire length of the structure. Before the outer jacket is applied, the areas over the threaded end molds must be filled and laminated with fiberglass roving and tape to the outside diameter of the copper rings, as is indicated in Fig. 4.

The region from the copper ring to the end of the structure forms a coupling. It includes a threaded section followed by a larger-diameter section which facilitates thread alignment when an adjoining section of copper pipe is inserted. This is followed by an end section still larger in diameter. The step formed by the difference in the two latter diameters provides a seat for a rubber "0" ring, which is slipped over the adjoining pipe before its entry and forced against the seat within the chamfered end section. The seal formed by this "0" ring permits evacuation of the assembled line and subsequent filling with dry nitrogen to avoid oxygen and moisture absorption losses.

Throughout the entire construction it is essential to maintain a uniform glass-to-resin content ratio and to apply a uniform tension to the glass cloth. Otherwise, the rigid dimensional requirements cannot be met. A non-uniform layering, commonly experienced with hand-layering, causes a non-uniform wall thickness. This results in non-uniform shrinkage upon curing, which sets up unequal internal stresses that are manifested by ellipticity and a warped axis. Considerable improvement in uniformity of layering has been accomplished by the

OUTER JACKET-LAMINATED TURNS OF GLASS CLOTH
INTERMEDIATE JACKET-LAMINATED TURNS OF GLASS CLOTH
LOSSY JACKET-LAMINATED TURNS OF TIN-OXIDE COATED GLASS CLOTH
INNER JACKET-LAMINATED TURNS OF GLASS CLOTH
HELIX-#37 FORMEX INSULATED COPPER WIRE CLOSE WOUND
COPPER RING
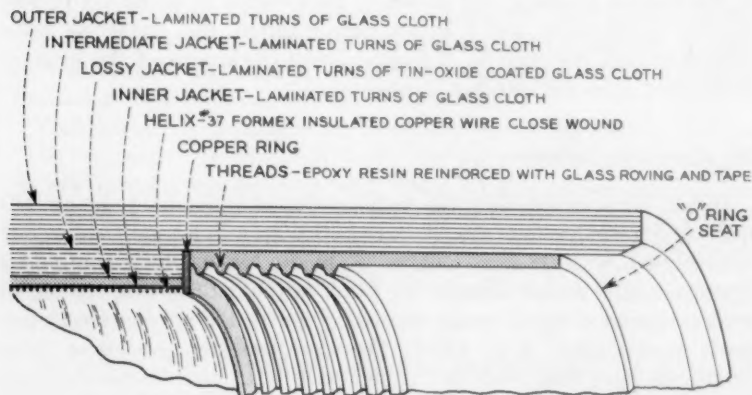THREADS-EPOXY RESIN REINFORCED WITH GLASS ROVING AND TAPE
"O" RING SEAT

Fig. 4 — Cross-sectional delineation of mode filter.

use of machine-layering, employing glass roving instead of glass cloth. By this means the glass content can be held uniformly at 75 per cent, while the hand-layered products showed 50 to 60 per cent glass content. The higher percentage of glass gives a stronger end product. To preserve a uniform wall thickness it was necessary to rotate the product continuously during layering and subsequent curing.

The curing cycle is also of very great importance in controlling the amount of ellipticity which develops in the finished product. Ellipticity is directly related to the amount of growth in the inside diameter as compared with the diameter of the mandrel on which the helix is wound. During curing, the resin shrinks toward the center of the plastic mass, which means it pulls away from the mandrel. The fiberglass reinforcing material, having a very low heat-expansion coefficient, tends not to move. Therefore, to minimize internal stresses, the shrinkage of the resin must be minimized. This can be done by curing it at as low a temperature as permissible over an extended period of time. During curing at low temperatures there is a sacrifice of heat-resisting qualities and tensile strength for the cured product. But the first of these qualities is of minor concern and the second is compensated for in the structural design.

After the product has been cured and has returned to room temperature, it is ready for the mandrel and thread molds to be extracted. The two ends are unscrewed and the mandrel is withdrawn with an hydraulic ram. The force required to extract the mandrel depends upon the amount of ellipticity in the finished product and the accuracy of the mandrel. The mandrel must meet dimensional requirements which preclude ridges in the surface grinding and present a high degree of polish. With low degrees of ellipticity and a satisfactory mandrel, the extracting force seldom exceeds 300 pounds. A slight amount of cleaning of the product is required after the mandrel is removed.

III. EXTENDED LENGTHS

It has been stated that the transmission characteristics of an all-helix-waveguide system should be superior to a solid-wall waveguide interspersed with 18-inch lengths of mode filters. Such continuous filtering would permit negotiating bends without distorting the signal with reconverted signal energy caused by conversion-reconversion intermode action (Ref. 5, p. 1230). Therefore, research models of helix waveguide have been made in 112-inch lengths. Construction techniques similar to those employed for the 18-inch lengths required using a special
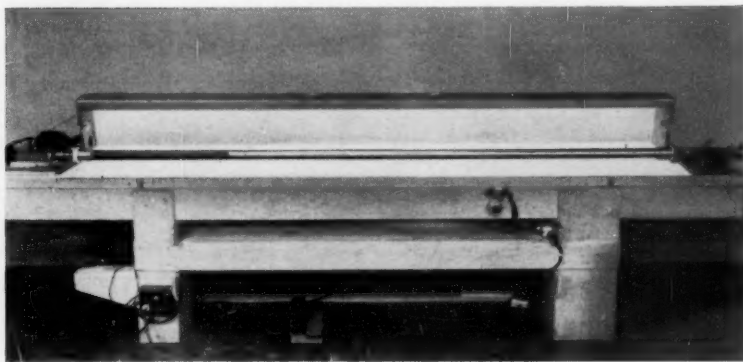
Fig. 5 — Mandrel and oven used in making 112-inch lengths of helix waveguide.

bench with a built-in curing oven which is illustrated in Fig. 5. Because of mechanical difficulties associated with hand-layering, the tolerances achieved on the 112-inch models did not equal those attained on the 18-inch models. With machine construction, it is expected that greater uniformity will be achieved on the longer helices.

IV. OVER-ALL CHARACTERISTICS OF THE HELICES

An inspection of a satisfactorily completed unit reveals the following performance characteristics:

4.1 *Visual*

Sighting through the unit reveals no apparent irregularities in the helical inner wall and, with proper illumination at the far end, one sees light diffraction exhibited as color rings. The outside surface shows a Barcol hardness indicator reading of 50 to 60.

Voids within the wall structure have been readily disclosed by X-rays. This has permitted improved technique. In a well constructed unit, X-rays reveal cross-sectional wall areas with void-free laminations. This type of examination verifies the fact that the number of voids is negligible in jackets constructed with machine-wound glass roving rather than with hand-layered glass cloth.

4.2 *Dimensional*

By means of an air gauge, maximum and minimum diameters are recorded at every inch of axial length, since it is the difference between
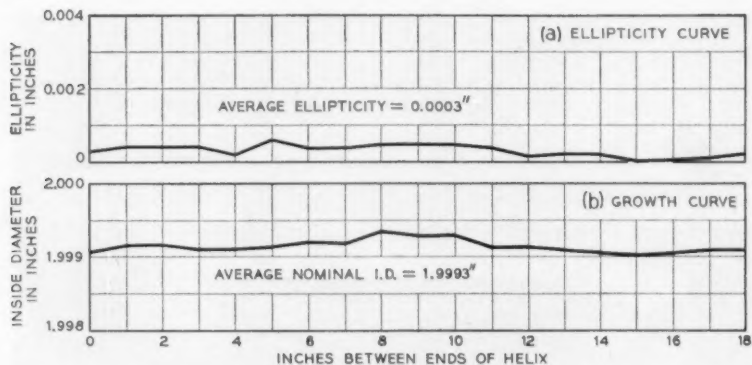
Fig. 6 — (a) Curve showing ellipticity in an 18-inch mode filter, (b) curve showing nominal inside diameters of an 18-inch mode filter.

the maximum and minimum diameters at any point along the axis that expresses the degree of ellipticity. A typical plot of this information is shown in Fig. 6(a). In Fig. 6(b) the average diameter computed from the above data is plotted as a function of position along the axis. The goal is a minimum degree of ellipticity and an average nominal inside diameter of 2.000 inches. For the 18-inch units, we are currently accepting degrees of ellipticity as herein defined which do not exceed 0.002 inch, with an average over-all ellipticity not greater than 0.001 inch. For the 112-inch units, the current degrees of ellipticity do not exceed 0.006 inch and the over-all average ellipticity is 0.0027 inch.

4.3 Bending Stiffness

When the mode filters are inserted at intervals in an otherwise solid-wall copper pipe line it is desirable to have their bending stiffness comparable, if not equal to, that of the copper pipe. Such structural uniformity throughout the line is necessary to insure that serpentine deformation caused by thermal expansion will be uniform in the entire line. If this were not so, a concentrated bend would occur at the weakest point, which would result in excessive axial tilt and introduce undesirable mode conversion. Therefore, the design of the mode filter is such that its resistance to bending, which is proportional to the product of its structural moment of inertia and its modulus of elasticity, must approximate that of the copper pipe. A jacket wall $\frac{5}{8}$ inch thick provides a sufficiently large structural moment of inertia to offset the greater modulus

of elasticity for the copper, which is 18×10^6 lbs. per sq. in. The modulus of elasticity of the reinforced plastic structure measures 2.5×10^6 lbs. per sq. in.

4.4 *Electrical*

The circular-electric-wave attenuation coefficient of the helices may be measured with appropriate apparatus which treats the unit as a resonant cavity. Such tests at 55 kmc* indicate that at present the best attenuation coefficient is 1.9 times that calculated for perfect solid-wall copper waveguide of the same diameter.

The other electrical performance characteristics of these helices will will be covered in another paper; the general problem of maximizing the loss for undesired modes of propagation is covered in detail by S. P. Morgan and J. A. Young.[7]

V. CONCLUSIONS

Sections of helix waveguide in 18- and 112-inch lengths have been made at the Holmdel Laboratory of Bell Telephone Laboratories. The shorter lengths of helix waveguide serve as mode filters to correct for conversion-reconversion effects in a transmission line composed mainly of solid-wall cylindrical waveguide. The longer helix units will provide continuous filtering when used exclusively in a helix-waveguide system. This becomes increasingly attractive for wide-band microwave transmission in the spectrum from 35,000 to 75,000 megacycles per second.

ACKNOWLEDGMENT

The author wishes to thank J. W. Bell and C. W. Curry for their help in the construction of these components.

REFERENCES

1. Schelkunoff, S. A., *Electromagnetic Waves*, D. Van Nostrand Company, Inc., New York, 1943.
2. Southworth, G. C., *Principles and Applications of Waveguide Transmission*, D. Van Nostrand Company, Inc., New York, 1950.

---

* The 55-kmc attenuation measurements were made by J. A. Young of the Laboratories.

3. Miller, S. E., Notes on Methods of Transmitting the Circular Electric Wave Around Bends, Proc. I.R.E., **40,** Sept., 1952, pp. 1104–113.
4. Miller, S. E. and Beck, A. C., Low-Loss Waveguide Transmission, Proc. I.R.E., **41,** March, 1953, pp. 348–358.
5. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., **33,** Nov., 1954, pp. 1209–1265.
6. King, A. P., Observed 5–6 MM Attenuation for the Circular Electric Wave in Small and Medium Sized Pipes, B.S.T.J. **35,** Sept., 1956, pp. 1115–1128.
7. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** Nov., 1956, pp. 1347–1384.

Table of First 700 Zeros of Bessel Functions — $J_l(x)$ and $J'_l(x)$

By CURTIS L. BEATTIE

(Manuscript received August 20, 1957)

The zeros of the Bessel functions and Bessel function deriviatives are identified by standard waveguide notation which also serves as a code for more general mathematical applications.

The possibilities of low-loss transmission using the TE_{01} (circular electric) mode in circular cylindrical pipe of a diameter large compared to the wavelength has made the study of other modes of such a waveguide important. In order to find phase and attenuation constants of various modes for both solid and ring or helix walls, the zeros of the Bessel functions $J_l(x)$ and $J'_l(x)$ are essential.

In the table given here the first seven hundred roots of Bessel functions $J_l(x) = 0$ and $J'_l(x) = 0$ have been computed and arranged in the order of the magnitude of the arguments corresponding to the roots. In the table l is the order of the Bessel function and m is the serial number of the zero of either $J_l(x)$ or $J'_l(x)$, not counting $x = 0$. In waveguide applications the zeros of $J_l(x)$ correspond to transverse magnetic modes of propagation (TM modes) and those of $J'_l(x)$ to transverse electric modes (TE modes). The designations TM and TE appear in the table for the benefit of those who will use this table in waveguide research and serve as a code designating $J_l(x)$ and $J'_l(x)$ for those who are interested in a more general application of the mathematics.

The roots of the Bessel functions were calculated from the *Tables of the Bessel Functions of the First Kind of Orders, J_0 through J_{51}*, computed by the Staff of the Computation Laboratory of Harvard University, published by the Harvard University Press, 1946–1948.

This table was first formulated horizontally in the ascending order of the function and vertically in the ascending number of the root. Since the increments in each direction are of a predictable magnitude, the possibility of having neglected a root is virtually eliminated.

All 700 roots were calculated by means of a linear interpolation for-

mula but checked and corrected for six-place accuracy for the first 300 roots. For arguments above 25, where the tabulated values were for each 0.01 only, the following Newton-Bessel formula was used:

$$J_n(x) = J_n(x_0 + h\mu) = J_n(x_0) + \mu B + \frac{\mu(\mu - 1)(C - A)}{4} + \cdots$$

where

$$\mu = (x - x_0)/h$$

$$A = J_n(x_0 + h) + J_n(x_0)$$

$$B = J_n(x_0 + h) - J_n(x_0)$$

$$C = J_n(x_0 + 2h) + J_n(x_0 - h).$$

Spot checking was done with the first three terms of the Taylor's series expansion:

$$0 \cong J_n(x) + \frac{h}{2}[J_{n-1}(x) - J_{n+1}(x)]$$

$$+ \frac{h^2}{8}[J_{n-2}(x) - 2J_n(x) + J_{n+2}(x)] + \cdots.$$

BIBLIOGRAPHY

1. Wilson, I. G., Schramm, C. W. and Kinzer, J. P., High Q Resonant Cavities for Microwave Testing, B.S.T.J., **25**, July, 1946, p. 418.
2. Gray, A. and Mathews, G. B., *Treatise on Bessel Functions*, Macmillan, London, 1895. Table III. The first 40 roots of $J_0(x) = 0$, with the corresponding values of $J_1(x)$. (Ten decimal places.) Table IV. The first 50 roots of $J_1(x) = 0$, with the corresponding maximum or minimum values of $J_0(x)$. (Sixteen decimal places.)
3. Smith, D. B., Rodgers, L. M. and Traub, E. H., Zeros of Bessel Functions, Franklin Inst. Jnl., **237**, April, 1944, p. 301. Positive real roots between 0 and 25 for function and first derivative of Bessel functions of the first kind.
4. Bickley, W. G., Notes on the Evaluation of Zeros and Turning Values of Bessel Functions, Interpolation by Taylor Series, Phil. Mag., 7th Series, **36**, March, 1945, p. 200; Miller, J. C. P. and Jones, C. W., Notes on the Evaluation of Zeros and Turning Values of Bessel Functions, Phil. Mag., 7th Series, **36**, March, 1945, p. 206.
5. Schelkunoff, S. A., *Electromagnetic Waves*, D. Van Nostrand and Co., N. Y., 1943.
6. Southworth, G. C., *Principles and Applications of Waveguide Transmission*, D. Van Nostrand and Co., N. Y., 1950.

TABLE

	Mode*	l_{-m}	Value†		Mode*	l_{-m}	Value†
1	TE	1–1	1.841184	(48	TM	1–4	13.323692
2	TM	0–1	2.404826	(49	TE	0–4	13.323692
3	TE	2–1	3.054237	50	TM	9–1	13.354300
(4	TM	1–1	3.831706	51	TM	6–2	13.589290
(5	TE	0–1	3.831706	52	TE	12–1	13.878843
6	TE	3–1	4.201189	53	TE	5–3	13.987189
7	TM	2–1	5.135622	54	TE	8–2	14.115519
8	TE	4–1	5.317553	55	TM	4–3	14.372537
9	TE	1–2	5.331443	56	TM	10–1	14.475501
10	TM	0–2	5.520078	57	TE	3–4	14.585848
11	TM	3–1	6.380162	58	TM	2–4	14.795952
12	TE	5–1	6.415616	59	TM	7–2	14.821269
13	TE	2–2	6.706133	60	TE	1–5	14.863589
(14	TM	1–2	7.015587	61	TE	13–1	14.928374
(15	TE	0–2	7.015587	62	TM	0–5	14.930918
16	TE	6–1	7.501266	63	TE	6–3	15.268181
17	TM	4–1	7.588342	64	TE	9–2	15.286738
18	TE	3–2	8.015237	65	TM	11–1	15.589848
19	TM	2–2	8.417244	66	TM	5–3	15.700174
20	TE	1–3	8.536316	67	TE	4–4	15.964107
21	TE	7–1	8.577836	68	TE	14–1	15.975439
22	TM	0–3	8.653728	69	TM	8–2	16.037774
23	TM	5–1	8.771484	70	TM	3–4	16.223466
24	TE	4–2	9.282396	71	TE	2–5	16.347522
25	TE	8–1	9.647422	72	TE	10–2	16.447853
26	TM	3–2	9.761023	(73	TM	1–5	16.470630
27	TM	6–1	9.936110	(74	TE	0–5	16.470630
28	TE	2–3	9.969468	75	TE	7–3	16.529366
(29	TM	1–3	10.173468	76	TM	12–1	16.698250
(30	TE	0–3	10.173468	77	TM	6–3	17.003820
31	TE	5–2	10.519861	78	TE	15–1	17.020323
32	TE	9–1	10.711434	79	TM	9–2	17.241220
33	TM	4–2	11.064709	80	TE	5–4	17.312842
34	TM	7–1	11.086370	81	TE	11–2	17.600267
35	TE	3–3	11.345924	82	TM	4–4	17.615966
36	TM	2–3	11.619841	83	TE	8–3	17.774012
37	TE	1–4	11.706005	84	TE	3–5	17.788748
38	TE	6–2	11.734936	85	TM	13–1	17.801435
39	TE	10–1	11.770877	86	TM	2–5	17.959819
40	TM	0–4	11.791534	87	TE	1–6	18.015528
41	TM	8–1	12.225092	88	TE	16–1	18.063261
42	TM	5–2	12.338604	89	TM	0–6	18.071064
43	TE	4–3	12.681908	90	TM	7–3	18.287583
44	TE	11–1	12.826491	91	TM	10–2	18.433464
45	TE	7–2	12.932386	92	TE	6–4	18.637443
46	TM	3–3	13.015201	93	TE	12–2	18.745091
47	TE	2–4	13.170371	94	TM	14–1	18.899998

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.

† $\bar{5}$ in last place indicates higher value and $\underline{5}$ indicates lower value in rounding off for fewer decimal places.

TABLE — *Continued*

	Mode* l_{-m}	Value†		Mode* l_{-m}	Value†
95	TM 5–4	18.980134	150	TM 4–6	24.019020
96	TE 9–3	19.004594	151	TE 3–7	24.144897
97	TE 17–1	19.104458	152	TM 9–4	24.233885
98	TE 4–5	19.196029	153	TM 15–2	24.269180
99	TM 3–5	19.409415	154	TM 2–7	24.270112
100	TE 2–6	19.512913	155	TE 22–1	24.289385
101	TM 8–3	19.554536	156	TE 1–8	24.311327
(102	TM 1–6	19.615859	157	TM 19–1	24.338250
(103	TE 0–6	19.615859	158	TM 0–8	24.352472
104	TM 11–2	19.615967	159	TE 17–2	24.381913
105	TE 13–2	19.883224	160	TM 12–3	24.494885
106	TE 7–4	19.941853	161	TE 8–5	24.587197
107	TM 15–1	19.994431	162	TM 7–5	24.934928
108	TE 18–1	20.144079	163	TE 14–3	25.001972
109	TE 10–3	20.223031	164	TE 11–4	25.008519
110	TM 6–4	20.320789	165	TE 6–6	25.183925
111	TE 5–5	20.575515	166	TE 23–1	25.322921
112	TM 12–2	20.789906	167	TM 16–2	25.417019
113	TM 9–3	20.807048	168	TM 20–1	25.417141
114	TM 4–5	20.826933	169	TM 5–6	25.430341
115	TE 3–6	20.972477	170	TE 18–2	25.495558
116	TE 14–2	21.015405	171	TM 10–4	25.509450
117	TM 16–1	21.085147	172	TE 4–7	25.589760
118	TM 2–6	21.116997	173	TM 13–3	25.705104
119	TE 1–7	21.164370	174	TM 3–7	25.748167
120	TE 19–1	21.182267	175	TE 2–8	25.826037
121	TM 0–7	21.211637	176	TE 9–5	25.891177
122	TE 8–4	21.229063	(177	TM 1–8	25.903672
123	TE 11–3	21.430854	(178	TE 0–8	25.903672
124	TM 7–4	21.641541	179	TE 15–3	26.177766
125	TE 6–5	21.931715	180	TE 12–4	26.246048
126	TM 13–2	21.956244	181	TM 8–5	26.266815
127	TM 10–3	22.046985	182	TE 24–1	26.355506
128	TE 15–2	22.142247	183	TM 21–1	26.493648
129	TM 17–1	22.172495	184	TE 7–6	26.545032
130	TM 5–5	22.217800	185	TM 17–2	26.559784
131	TE 20–1	22.219145	186	TE 19–2	26.605533
132	TE 4–6	22.401032	187	TM 11–4	26.773323
133	TE 9–4	22.501399	188	TM 6–6	26.820152
134	TM 3–6	22.582730	189	TM 14–3	26.907369
135	TE 12–3	22.629300	190	TE 5–7	27.010308
136	TE 2–7	22.671582	191	TE 10–5	27.182022
(137	TM 1–7	22.760084	192	TM 4–7	27.199088
(138	TE 0–7	22.760084	193	TE 3–8	27.310058
139	TM 8–4	22.945173	194	TE 16–3	27.347386
140	TM 14–2	23.115778	195	TE 25–1	27.387204
141	TE 21–1	23.254816	196	TM 2–8	27.420574
142	TM 18–1	23.256777	197	TE 1–9	27.457051
143	TE 16–2	23.264269	198	TE 13–4	27.474340
144	TE 7–5	23.268053	199	TM 0–9	27.493480
145	TM 11–3	23.275854	200	TM 22–1	27.567944
146	TM 6–5	23.586084	201	TM 9–5	27.583749
147	TE 10–4	23.760716	202	TM 18–2	27.697899
148	TE 5–6	23.803581	203	TE 20–2	27.712126
149	TE 13–3	23.819374	204	TE 8–6	27.889270

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.
    † $\bar{5}$ in last place indicates higher value and $\underline{5}$ indicates lower value in rounding off for fewer decimal places.

Table — *Continued*

	Mode* l–m	Value†		Mode* l–m	Value†
205	TM 12–4	28.026710	256	TM 12–5	31.459960
206	TM 15–3	28.102416	257	TE 29–1	31.506195
207	TM 7–6	28.191189	258	TE 6–8	31.617876
208	TE 6–7	28.409776	259	TM 18–3	31.650118
209	TE 26–1	28.418072	260	TM 15–4	31.733414
210	TE 11–5	28.460857	261	TM 5–8	31.811717
211	TE 17–3	28.511361	262	TE 11–6	31.838425
212	TM 5–7	28.626619	263	TM 26–1	31.845888
213	TM 23–1	28.640185	264	TE 4–9	31.938540
214	TE 14–4	28.694271	265	TE 20–3	31.973715
215	TE 4–8	28.767836	266	TM 3–9	32.064853
216	TE 21–2	28.815590	267	TE 24–2	32.109320
217	TM 19–2	28.831731	268	TE 2–10	32.127327
218	TM 10–5	28.887375	(269	TM 1–10	32.189680
219	TM 3–8	28.908351	(270	TE 0–10	32.189680
220	TE 2–9	28.977673	271	TM 22–2	32.210587
(221	TM 1–9	29.046829	272	TM 10–6	32.211856
(222	TE 0–9	29.046829	273	TE 14–5	32.236970
223	TE 9–6	29.218564	274	TE 17–4	32.310894
224	TM 13–4	29.270631	275	TE 9–7	32.505248
225	TM 16–3	29.290871	276	TE 30–1	32.534220
226	TE 27–1	29.448163	277	TM 13–5	32.731053
227	TM 8–6	29.545660	278	TM 8–7	32.795800
228	TE 18–3	29.670147	279	TM 19–3	32.821803
229	TM 24–1	29.710509	280	TM 27–1	32.911154
230	TE 12–5	29.728978	281	TM 16–4	32.953665
231	TE 7–7	29.790749	282	TE 7–8	33.015179
232	TE 15–4	29.906591	283	TE 21–3	33.119162
233	TE 22–2	29.916147	284	TE 12–6	33.131450
234	TM 20–2	29.961604	285	TE 25–2	33.202272
235	TM 6–7	30.033723	286	TM 6–8	33.233042
236	TM 11–5	30.179061	287	TM 23–2	33.330177
237	TE 5–8	30.202849	288	TE 5–9	33.385444
238	TM 4–8	30.371008	289	TE 15–5	33.478449
239	TE 3–9	30.470269	290	TE 18–4	33.503929
240	TM 17–3	30.473280	291	TM 11–6	33.526364
241	TE 28–1	30.477523	292	TM 4–9	33.537138
242	TM 14–4	30.505951	293	TE 31–1	33.561634
243	TE 10–6	30.534505	294	TE 3–10	33.626949
244	TM 2–9	30.569205	295	TM 2–10	33.716520
245	TE 1–10	30.601923	296	TE 1–11	33.746183
246	TM 0–10	30.634607	297	TM 0–11	33.775821
247	TM 25–1	30.779039	298	TE 10–7	33.841966
248	TE 19–3	30.824148	299	TM 28–1	33.974930
249	TM 9–6	30.885379	300	TM 20–3	33.988703
250	TE 13–5	30.987394	301	TM 14–5	33.99319
251	TE 23–2	31.013998	302	TM 9–7	34.15438
252	TM 21–2	31.087805	303	TM 17–4	34.16727
253	TE 16–4	31.111945	304	TE 22–3	34.26077
254	TE 8–7	31.155327	305	TE 26–2	34.29300
255	TM 7–7	31.422795	306	TE 8–8	34.39663

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.

† $\bar{5}$ in last place indicates higher value and 5 indicates lower value in rounding off for fewer decimal places.

TABLE — *Continued*

	Mode* *l–m*	Value†		Mode* *l–m*	Value†
307	TE 13–6	34.41455	357	TE 10–8	37.11800
308	TM 24–2	34.44678	358	TM 31–1	37.15811
309	TE 32–1	34.58847	359	TE 18–5	37.16040
310	TM 7–8	34.63709	360	TM 9–8	37.40010
311	TE 19–4	34.69148	361	TM 14–6	37.40819
312	TE 16–5	34.71248	362	TM 23–3	37.46381
313	TE 6–9	34.81339	363	TE 29–2	37.55307
314	TM 12–6	34.82999	364	TE 8–9	37.62008
315	TM 5–9	34.98878	365	TE 25–3	37.66491
316	TM 29–1	35.03730	366	TE 35–1	37.66577
317	TE 4–10	35.10392	367	TM 17–5	37.73268
318	TM 21–3	35.15115	368	TM 20–4	37.77286
319	TE 11–7	35.16671	369	TM 27–2	37.78040
320	TM 3–10	35.21867	370	TE 13–7	37.78438
321	TM 15–5	35.24709	371	TM 7–9	37.83872
322	TE 2–11	35.27554	372	TE 6–10	37.99964
(323	TM 1–11	35.33231	373	TM 12–7	38.15638
(324	TE 0–11	35.33231	374	TM 5–10	38.15987
325	TM 18–4	35.37472	375	TE 16–6	38.21206
326	TE 27–2	35.38163	376	TM 32–1	38.21669
327	TE 23–3	35.39878	377	TE 22–4	38.22490
328	TM 10–7	35.49991	378	TE 4–11	38.26532
329	TM 25–2	35.56057	379	TM 3–11	38.37047
330	TE 33–1	35.61475	380	TE 19–5	38.37524
331	TE 14–6	35.68854	381	TE 2–12	38.42266
332	TE 9–8	35.76379	382	TE 11–8	38.46039
333	TE 20–4	35.87394	(383	TM 1–12	38.47477
334	TE 17–5	35.93963	(384	TE 0–12	38.47477
335	TM 8–8	36.02562	385	TM 24–3	38.61452
336	TM 30–1	36.09834	386	TE 30–2	38.63609
337	TM 13–6	36.12366	387	TM 15–6	38.68428
338	TE 7–9	36.22438	388	TE 36–1	38.69055
339	TM 22–3	36.30943	389	TM 10–8	38.76181
340	TM 6–9	36.42202	390	TE 26–3	38.79341
341	TE 28–2	36.46829	391	TM 28–2	38.88671
342	TE 12–7	36.48055	392	TM 21–4	38.96429
343	TM 16–5	36.49340	393	TM 18–5	38.96543
344	TE 24–3	36.53343	394	TE 9–9	39.00190
345	TE 5–10	36.56078	395	TE 14–7	39.07900
346	TM 19–4	36.57645	396	TM 8–9	39.24045
347	TE 34–1	36.64051	397	TM 33–1	39.27413
348	TM 26–2	36.67173	398	TE 23–4	39.39398
349	TM 4–10	36.69900	399	TE 7–10	39.42227
350	TE 3–11	36.78102	400	TE 17–6	39.46277
351	TM 11–7	36.83357	401	TM 13–7	39.46921
352	TM 2–11	36.86286	402	TE 20–5	39.58453
353	TE 1–12	36.88999	403	TM 6–10	39.60324
354	TM 0–12	36.91710	404	TE 37–1	39.71489
355	TE 15–6	36.95417	405	TE 31–2	39.71743
356	TE 21–4	37.05164	406	TE 5–11	39.73064

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.

† 5 in last place indicates higher value and 5̲ indicates lower value in rounding off for fewer decimal places.

TABLE — *Continued*

	Mode* L_m	Value†		Mode* L_m	Value†
407	TM 25–3	39.76179	458	TE 29–3	42.16260
408	TE 12–8	39.79194	459	TM 31–2	42.19275
409	TM 4–11	39.85763	460	TE 9–10	42.22464
410	TE 27–3	39.91909	461	TE 14–8	42.42585
411	TE 3–12	39.93311	462	TM 36–1	42.44014
412	TM 16–6	39.95255	463	TM 8–10	42.44389
413	TM 29–2	39.99080	464	TM 18–6	42.46781
414	TM 2–12	40.00845	465	TM 24–4	42.51168
415	TE 1–13	40.03344	466	TE 7–11	42.61152
416	TM 0–13	40.05843	467	TM 21–5	42.62870
417	TM 11–8	40.11182	468	TM 6–11	42.77848
418	TM 22–4	40.15105	469	TM 13–8	42.78044
419	TM 19–5	40.19210	470	TE 40–1	42.78537
420	TM 34–1	40.33048	471	TE 26–4	42.87855
421	TE 15–7	40.36510	472	TE 5–12	42.89627
422	TE 10–9	40.37107	473	TE 17–7	42.91415
423	TE 24–4	40.55913	474	TE 34–2	42.95218
424	TM 9–9	40.62855	475	TM 4–12	43.01374
425	TE 18–6	40.70680	476	TE 12–9	43.07549
426	TE 38–1	40.73879	477	TE 3–13	43.08365
427	TM 14–7	40.77283	478	TM 2–13	43.15345
428	TE 21–5	40.78864	479	TE 20–6	43.17654
429	TE 32–2	40.79718	480	TE 1–14	43.17663
430	TE 8–10	40.83018	481	TE 23–5	43.18255
431	TM 26–3	40.90580	482	TM 28–3	43.18477
432	TM 7–10	41.03077	483	TM 0–14	43.19979
433	TE 28–3	41.04211	484	TE 30–3	43.28071
434	TM 30–2	41.09278	485	TM 32–2	43.29081
435	TE 13–8	41.11351	486	TM 16–7	43.35507
436	TE 6–11	41.17885	487	TM 11–9	43.36836
437	TM 17–6	41.21357	488	TM 37–1	43.49352
438	TM 5–11	41.32638	489	TE 10–10	43.60677
439	TM 23–4	41.33343	490	TM 25–4	43.68603
440	TM 35–1	41.38580	491	TM 19–6	43.71571
441	TM 20–5	41.41307	492	TE 15–8	43.72963
442	TE 4–12	41.42367	493	TE 41–1	43.80808
443	TM 12–8	41.45109	494	TM 22–5	43.83932
444	TM 3–12	41.52072	495	TM 9–10	43.84380
445	TE 2–13	41.56894	496	TE 35–2	44.02758
(446	TM 1–13	41.61709	497	TE 8–11	44.03001
(447	TE 0–13	41.61709	498	TE 27–4	44.03321
448	TE 16–7	41.64331	499	TM 14–8	44.10059
449	TE 25–4	41.72059	500	TE 18–7	44.17813
450	TE 11–9	41.72863	501	TM 7–11	44.21541
451	TE 39–1	41.76228	502	TM 29–3	44.32003
452	TE 33–2	41.87540	503	TE 6–12	44.35258
453	TE 19–6	41.94459	504	TE 24–5	44.37290
454	TE 22–5	41.98788	505	TM 33–2	44.38706
455	TM 10–9	42.00419	506	TE 31–3	44.39653
456	TM 27–3	42.04674	507	TE 21–6	44.40300
457	TM 15–7	42.06792	508	TE 13–9	44.41245

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.

† 5 in last place indicates higher value and 5̲ indicates lower value in rounding off for fewer decimal places.

TABLE — *Continued*

	Mode* l_m	Value†		Mode* l_m	Value†
509	TM 5–12	44.48932	561	TE 20–7	46.68717
510	TM 38–1	44.54601	562	TM 16–8	46.71581
511	TE 4–13	44.57962	563	TE 26–5	46.74158
512	TM 17–7	44.63483	564	TE 10–11	48.82896
513	TM 3–13	44.66974	565	TE 23–6	46.84075
514	TE 2–14	44.71455	566	TE 44–1	46.87409
515	TM 12–9	44.72194	567	TM 9–11	47.04870
(516	TM 1–14	44.75932	568	TE 15–9	47.05946
(517	TE 0–14	44.75932	569	TM 19–7	47.17400
518	TE 42–1	44.83043	570	TM 28–4	47.18775
519	TM 26–4	44.85670	571	TE 8–12	47.22176
520	TM 20–6	44.95768	572	TE 38–2	47.24608
521	TE 11–10	44.97753	573	TM 7–12	47.39417
522	TE 16–8	45.02543	574	TM 14–9	47.40035
523	TM 23–5	45.04521	575	TM 22–6	47.42517
524	TE 36–2	45.10166	576	TM 25–5	47.44385
525	TE 28–4	45.18473	577	TE 30–4	47.47899
526	TM 10–10	45.23157	578	TE 6–13	47.52196
527	TM 15–8	45.41219	579	TE 18–8	47.59513
528	TE 9–11	45.43548	580	TM 5–13	47.64940
529	TE 19–7	45.43567	581	TM 36–2	47.66568
530	TM 30–3	45.45267	582	TE 13–10	47.68825
531	TM 34–2	45.48156	583	TM 41–1	47.69840
532	TE 32–3	45.51018	584	TM 32–3	47.71055
533	TE 25–5	45.55917	585	TE 34–3	47.73138
534	TM 39–1	45.59762	586	TE 4–14	47.73367
535	TE 22–6	45.62431	587	TM 3–14	47.81779
536	TM 8–11	45.63844	588	TE 2–15	47.85964
537	TE 14–9	45.74024	589	TE 45–1	47.89542
538	TE 7–12	45.79400	(590	TM 1–15	47.90146
539	TE 43–1	45.85243	(591	TE 0–15	47.90146
540	TM 18–7	45.90766	592	TE 27–5	47.92033
541	TM 6–12	45.94902	593	TE 21–7	47.93298
542	TM 27–4	46.02388	594	TM 12–10	47.97429
543	TE 5–13	46.05857	595	TM 17–8	48.01196
544	TM 13–9	46.06571	596	TE 24–6	48.05260
545	TM 4–13	46.16785	597	TE 11–11	48.21133
546	TE 37–2	46.17447	598	TE 39–2	48.31652
547	TM 21–6	46.19406	599	TM 29–4	48.34846
548	TE 3–14	46.23297	600	TE 16–9	48.37069
549	TM 24–5	46.24664	601	TM 20–7	48.43424
550	TM 2–14	46.29800	602	TM 10–11	48.44715
551	TE 17–8	46.31377	603	TE 31–4	48.62201
552	TE 1–15	46.31960	604	TE 9–12	48.63692
553	TE 29–4	46.33328	605	TM 26–5	48.63706
554	TE 12–10	46.33777	606	TM 23–6	48.65132
555	TM 0–15	46.34119	607	TM 15–9	48.72646
556	TM 35–2	46.57441	608	TM 42–1	48.74762
557	TM 31–3	46.58280	609	TM 37–2	48.75542
558	TM 11–10	46.60813	610	TM 8–12	48.82593
559	TE 33–3	46.62177	611	TM 33–3	48.83603
560	TM 40–1	46.64841	612	TE 35–3	48.83910

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.
† 5 in last place indicates higher value and 5 indicates lower value in rounding off for fewer decimal places.

TABLE—*Concluded*

	Mode* L_m	Value†		Mode* L_m	Value†
613	TE 19–8	48.86993	657	TE 6–14	50.68782
614	TE 46–1	48.91645	658	TM 5–14	50.80717
615	TE 7–13	48.97107	659	TM 44–1	50.84387
616	TE 14–10	49.02964	660	TE 4–15	50.88616
617	TE 28–5	49.09560	661	TE 33–4	50.90045
618	TM 6–13	49.11577	662	TM 39–2	50.93060
619	TE 22–7	49.17342	663	TE 48–1	50.93760
620	TE 5–14	49.21817	664	TM 22–7	50.93776
621	TE 25–6	49.26009	665	TE 13–11	50.94585
622	TM 18–8	49.30111	666	TM 3–15	50.96503
623	TM 4–14	49.32036	667	TE 18–9	50.97113
624	TM 13–10	49.33078	668	TE 2–16	51.00430
625	TE 3–15	49.38130	669	TM 28–5	51.01228
626	TE 40–2	49.38586	(670	TM 1–16	51.04354
627	TM 2–15	49.44216	(671	TE 0–16	51.04354
628	TE 1–16	49.46239	672	TE 37–3	51.04919
629	TM 0–16	49.48261	673	TM 35–3	51.08055
630	TM 30–4	49.50618	674	TM 25–6	51.08975
631	TE 12–11	49.58340	675	TM 12–11	51.21197
632	TE 17–9	49.67443	676	TM 17–9	51.35527
633	TM 21–7	49.68872	677	TE 21–8	51.40137
634	TE 32–4	49.76246	678	TE 11–12	51.43311
635	TM 43–1	49.79610	679	TE 30–5	51.43637
636	TM 77–5	49.82648	680	TE 42–2	51.52135
637	TM 11–11	49.83465	681	TE 24–7	51.63937
638	TM 38–2	49.84371	682	TM 10–12	51.65325
639	TM 24–6	49.87276	683	TE 27–6	51.66288
640	TE 47–1	49.93717	684	TE 16–10	51.68742
641	TE 36–3	49.94501	685	TM 32–4	51.81316
642	TM 34–3	49.95933	686	TE 9–13	51.83078
643	TE 10–12	50.04043	687	TM 20–8	51.86002
644	TM 16–9	50.04461	688	TM 45–1	51.89095
645	TE 20–8	50.13856	689	TE 49–1	51.97776
646	TM 9–12	50.24533	690	TM 8–13	52.00769
647	TE 29–5	50.26756	691	TM 40–2	52.01615
648	TE 15–10	50.36251	692	TM 15–10	52.01724
649	TE 8–13	50.40702	693	TE 34–4	52.03608
650	TE 23–7	50.40880	694	TE 7–14	52.14375
651	TE 41–2	50.45412	695	TE 38–3	52.15171
652	TE 26–6	50.46345	696	TM 23–7	52.18166
653	TM 7–13	50.56818	697	TM 29–5	52.19465
654	TM 19–8	50.58367	698	TM 36–3	52.19978
655	TM 31–4	50.66103	699	TE 19–9	52.26121
656	TM 14–10	50.67824	700	TM 6–14	52.27945

* TM designates a zero of $J_l(x)$; TE designates a zero of $J'_l(x)$; in each case l corresponds to the order of the Bessel function and m is the number of the root.

† 5 in last place indicates higher value and 5 indicates lower value in rounding off for fewer decimal places.
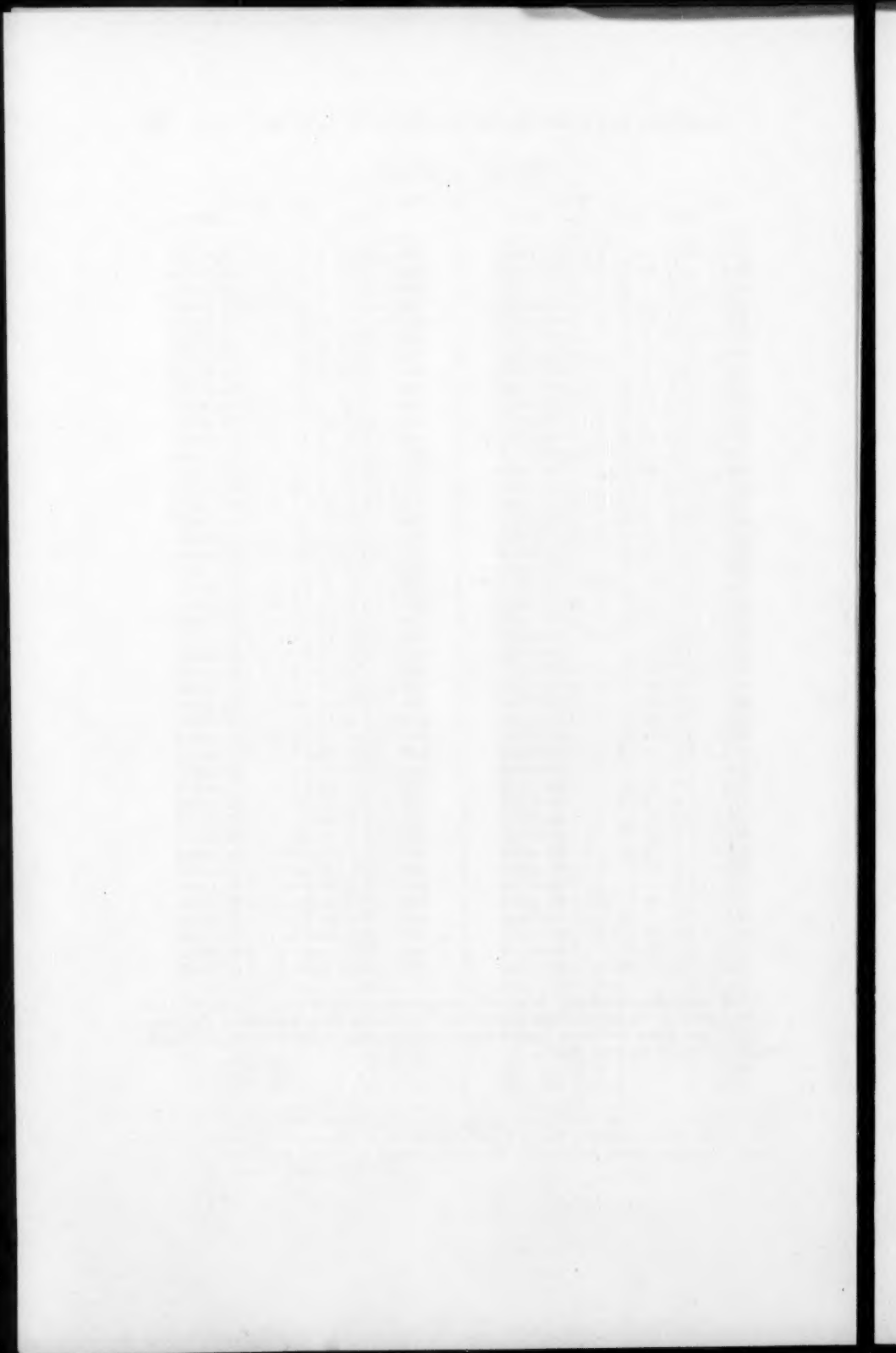
Evaluation of the Surface Concentration of Diffused Layers in Silicon

By GERHARD BACKENSTOSS

A method for determining the surface concentration of diffused impurity layers in semiconductors is described. It is shown that the surface concentration may be evaluated if the sheet resistivity of the layer, junction depth, impurity distribution across the layer, background resistivity and the majority carrier mobility are known. The calculations have been made for several typical impurity distributions in silicon.

I. INTRODUCTION

Recently the investigation of the diffusion of electrically active impurities in semiconducting materials has become important because of the use of diffusion technology in the fabrication of semiconductor devices. A diffused layer is characterized by the impurity distribution; its mathematical form is obtained by solution of the diffusion equation with the appropriate boundary conditions. When the distribution is known, two of the following parameters may be chosen to describe the layer completely: diffusion coefficient — time product, surface concentration, total number of diffused atoms and concentration at a given depth. Of these quantities, only the depth is easily measured. Another readily measured quantity of the layer is its sheet resistivity. But, from the point of view of the device designer, the junction depth and surface concentration are the most important quantities. The present paper shows how the surface concentration of a diffused layer may be obtained from the sheet resistivity and junction depth if the impurity distribution, the majority carrier mobilities, and the resistivity of the silicon are known. The necessary evaluations are given here for the case of diffusion of Group III and V elements into uniformly doped silicon of the opposite conductivity type.
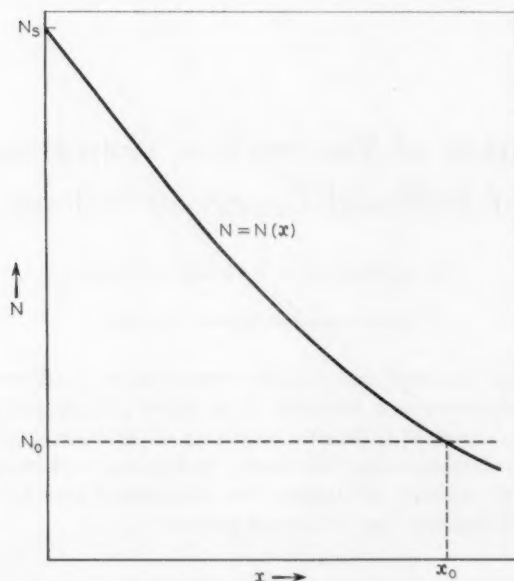
Fig. 1 — Assumed impurity distribution, with N_0 representing the impurity concentration of the original material.

II. METHOD

Fig. 1 shows an assumed impurity distribution, with N_0 representing the impurity concentration of the original material. A p-n junction is formed at a depth x_0 where the impurity concentration of the diffusant equals N_0. The depth of this junction x_0 can be obtained by one of several techniques.[1, 2] Thus a concentration at a given depth is known. If all impurities are ionized, the following relation holds for the average conductivity $\bar{\sigma}$ of an impurity layer of the thickness x_0,

$$\bar{\sigma} = \frac{1}{\rho_s x_0} = \frac{1}{x_0} q \int_0^{x_0} \mu(n)[N(x) - N_0] \, dx, \qquad (1)$$

with the conditions

$$N(x_0) = N_0, \qquad (2)$$

and $N(0) = N_s$ where N_s is the surface concentration. The quantity ρ_s can be obtained by a four-point probe measurement[3] on the surface. In silicon the saturation currents are so low that the junction may be regarded as an open circuit, so that the measured sheet resistivity is
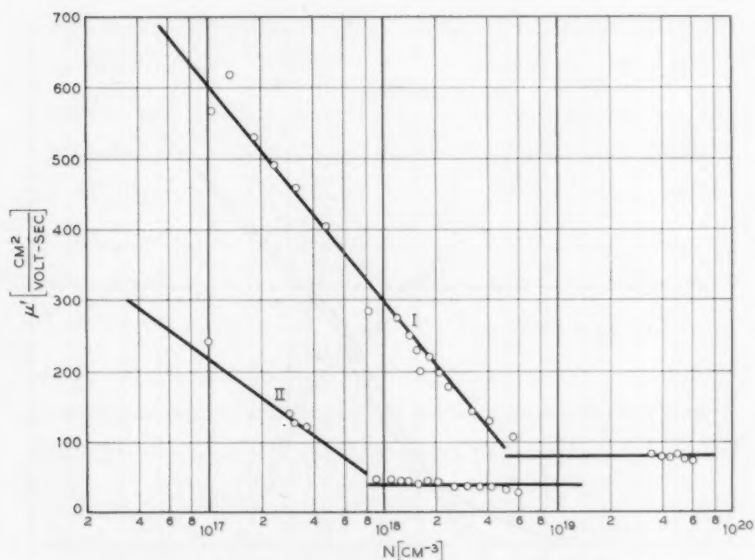
Fig. 2 — Effective mobility $\mu'(N)$ as determined by radioactive techniques for donor (curve I), and acceptor (curve II) elements in silicon.

essentially that of the layer alone. $N(x)$ represents the impurity density as a function of the distance x from the surface, and μ is the majority carrier conductivity mobility, which depends on the ionized impurity density n. Since only a fraction of the impurity atoms are ionized, particularly in heavily doped silicon, an appropriate correction must be made. This fact may be taken into account more directly by introducing and effective mobility $\mu'(N)$ defined by:

$$\mu' = 1/q\rho N, \tag{3}$$

where q is the magnitude of the electron charge and ρ the resistivity. The effective mobility $\mu'(N)$ has been determined by radioactive techniques for different impurity elements[4] and is shown in Fig. 2.

The results of these measurements can be described approximately by an analytic expression of the form

$$\mu_n' = 300 \ (19 - \log N) \ \text{for} \ 10^{16} < N < 5.10^{18} \ \text{cm}^{-3}$$

$$\mu_n' = 80 \qquad\qquad\quad \text{for} \ 5.10^{18} < N < 10^{20} \ \text{cm}^{-3} \tag{4a}$$

for electrons, and
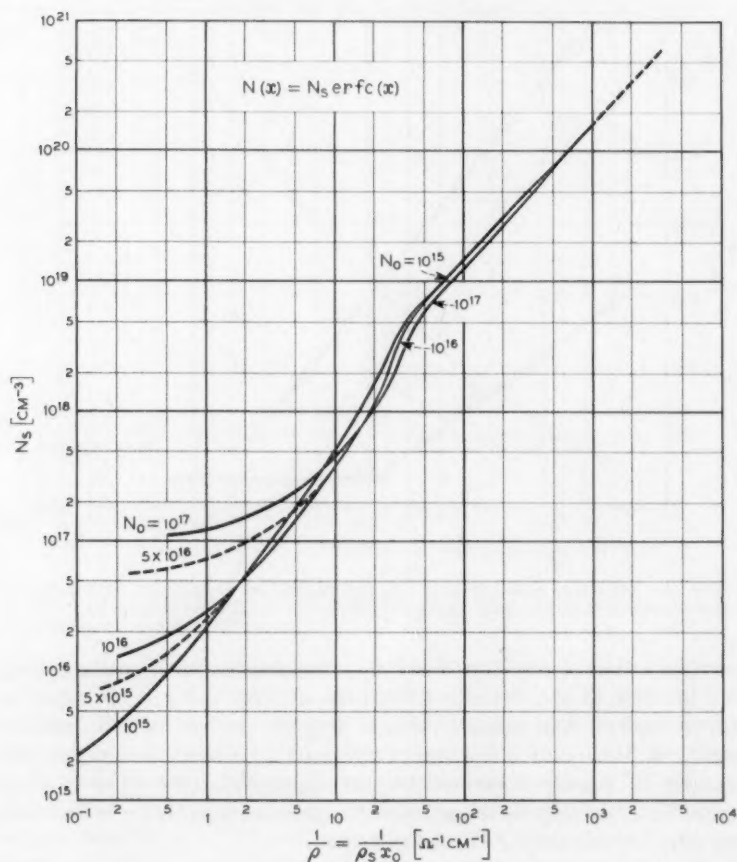
Fig. 3 — Surface concentration N_s versus conductivity $(1/\rho)$ for an n-type layer; complementary error function distribution.

$$\mu_p' = 180 \ (18.2 - \log N) \text{ for } 10^{16} < N < 8.10^{17} \text{ cm}^{-3}$$

$$\mu_p' = 40 \qquad\qquad \text{for } 8.10^{17} < N < 10^{20} \text{ cm}^{-3} \tag{4b}$$

for holes.

The hole mobility μ_p' was measured only to 6.10^{18} cm^{-3} but its decrease for higher concentrations is so small that the assumption of a constant mobility up to 10^{20} cm^{-3} is justified. With a specified distribution, the
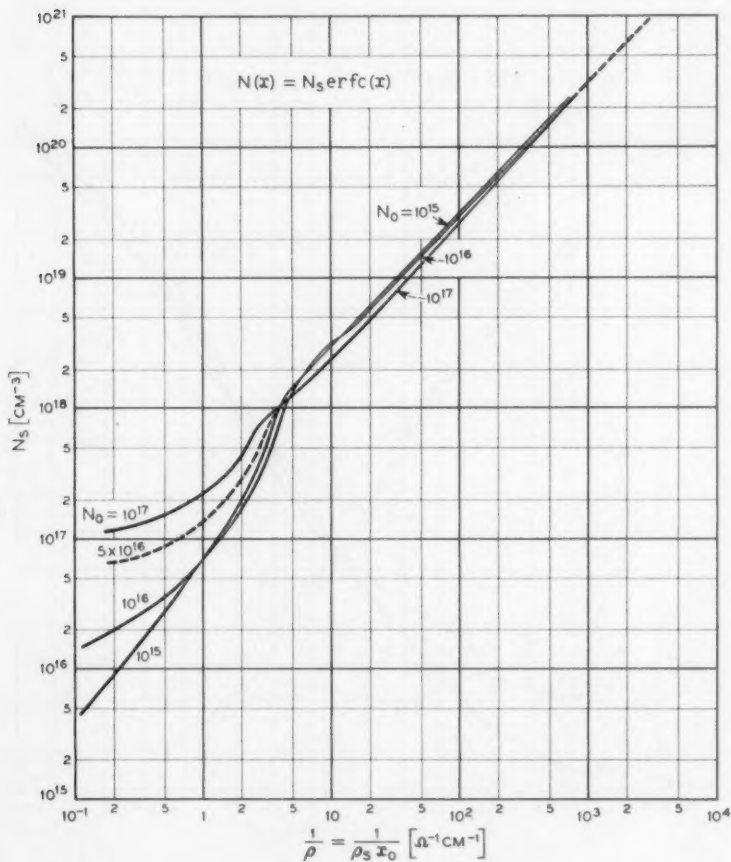
Fig. 4 — Surface concentration N_s versus conductivity $(1/\rho)$ for a p-type layer; complementary error function distribution.

surface concentration, N_s, can be evaluated as a function of $1/\rho_s x_0$ from (1), (2) and (4).

III. RESULTS

The calculations have been carried out for several distinct distributions which occur in different diffusion processes.
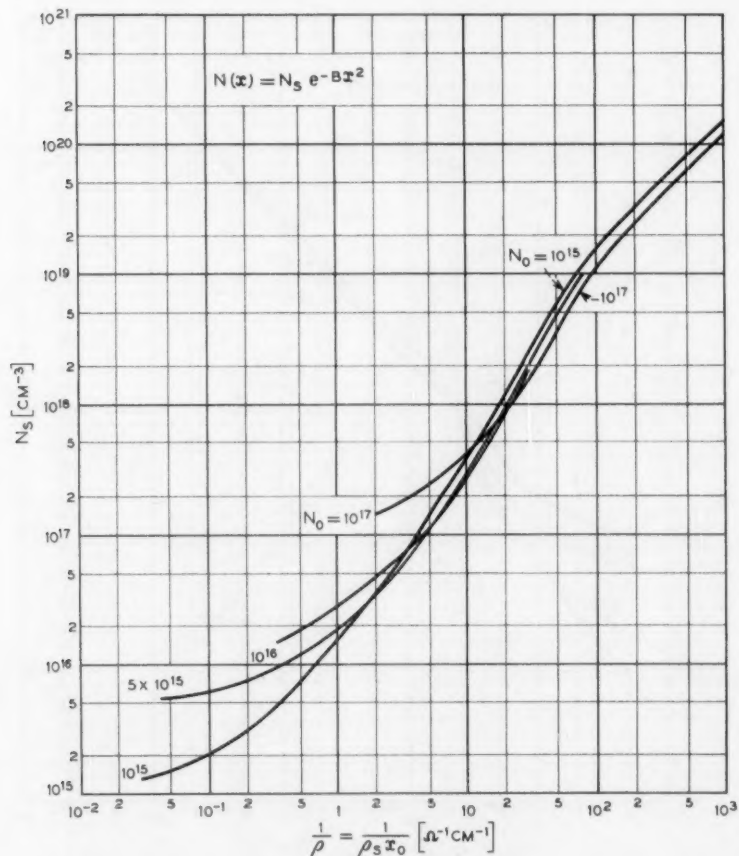
Fig. 5 — Surface concentration N_s versus conductivity for an n-type layer; Gaussian distribution.

Case I:

The maintenance of a constant impurity concentration at the surface during the diffusion leads to the complementary error function solution of the diffusion equation, that is

$$N(x) = N_s \, \text{erfc} \, (x/2\sqrt{Dt}).$$

The results are plotted in Figs. 3 and 4 for an n-type and a p-type layer
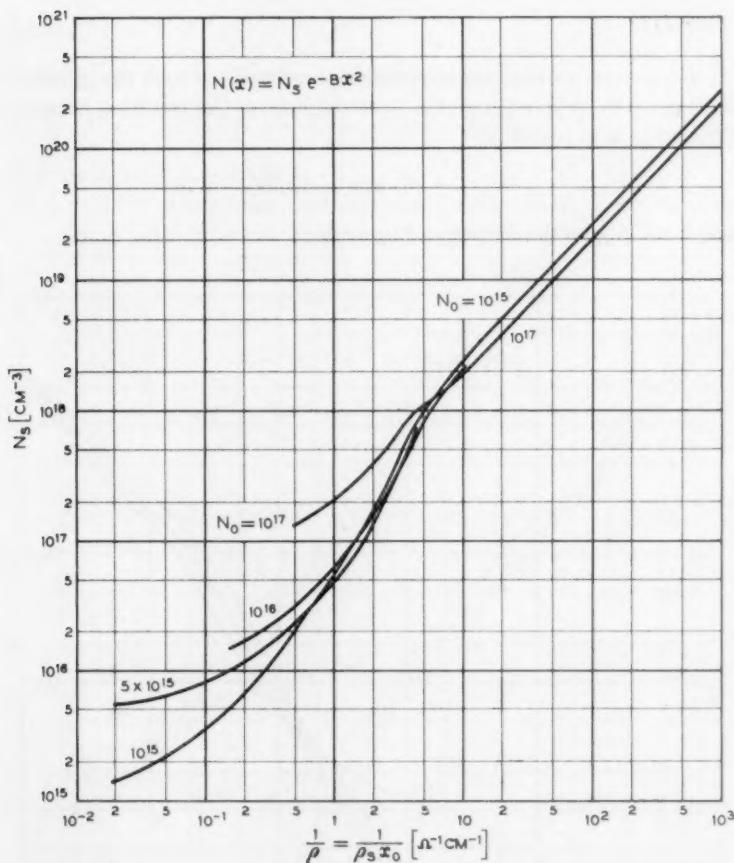
Fig. 6 — Surface concentration N_s versus conductivity for a p-type layer; Gaussian distribution.

respectively. N_s is plotted versus $1/\rho_s x_0$ with background concentrations N_0 as a parameter.

Case II:

If the initial impurity distribution is represented by a δ-function, one obtains after diffusion a Gaussian distribution

$$N(x) = N_s \exp\left(-x^2/4Dt\right).$$

Figs. 5 and 6 show corresponding plots for n- and p-type layers.

Case III:

If a constant surface concentration is maintained and the silicon is simultaneously evaporating at a constant rate, v, the resulting impurity distribution is exponential[5]

$$N(x) = N_s \exp(-vx/D).$$

Figs. 7 and 8 give the corresponding plots.
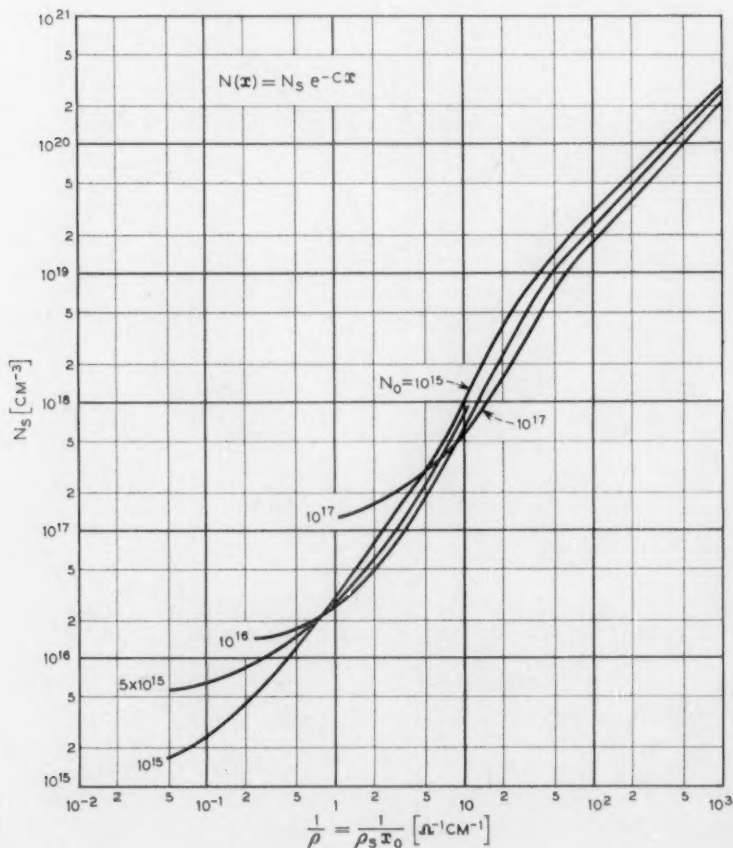
Fig. 7 — Surface concentration N_s versus conductivity for an n-type layer; exponential distribution.

Case IV:

Finally, curves for a linear impurity distribution

$$N(x) = N_s - (N_s - N_0)x/x_0$$

are given in Figs. 9 and 10.

It should be pointed out that the effective mobilities μ' are different for elements with different ionization energies. However, the ionization
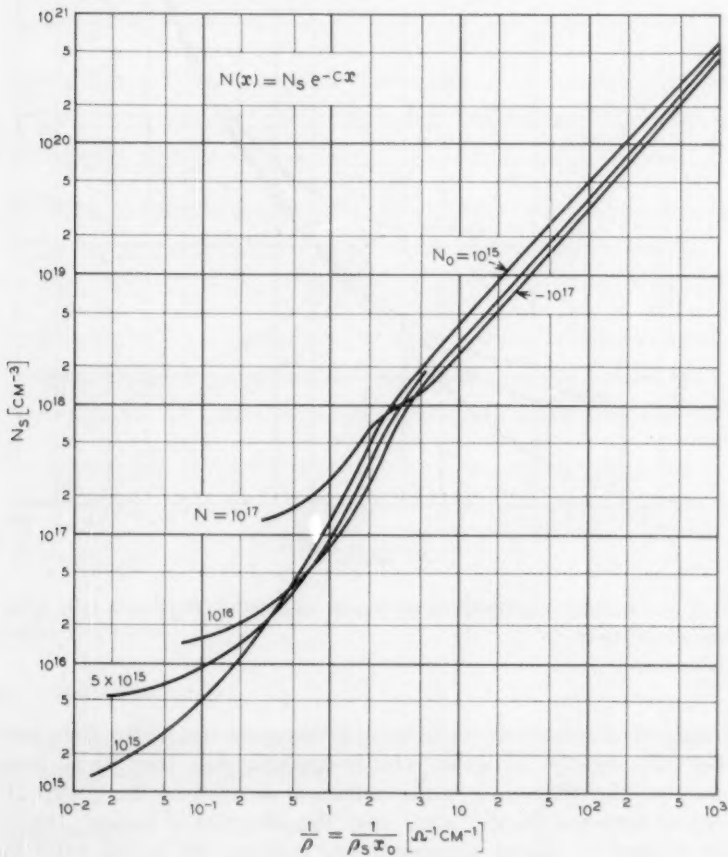
Fig. 8 — Surface concentration N_s versus conductivity for a p-type layer; exponential distribution.
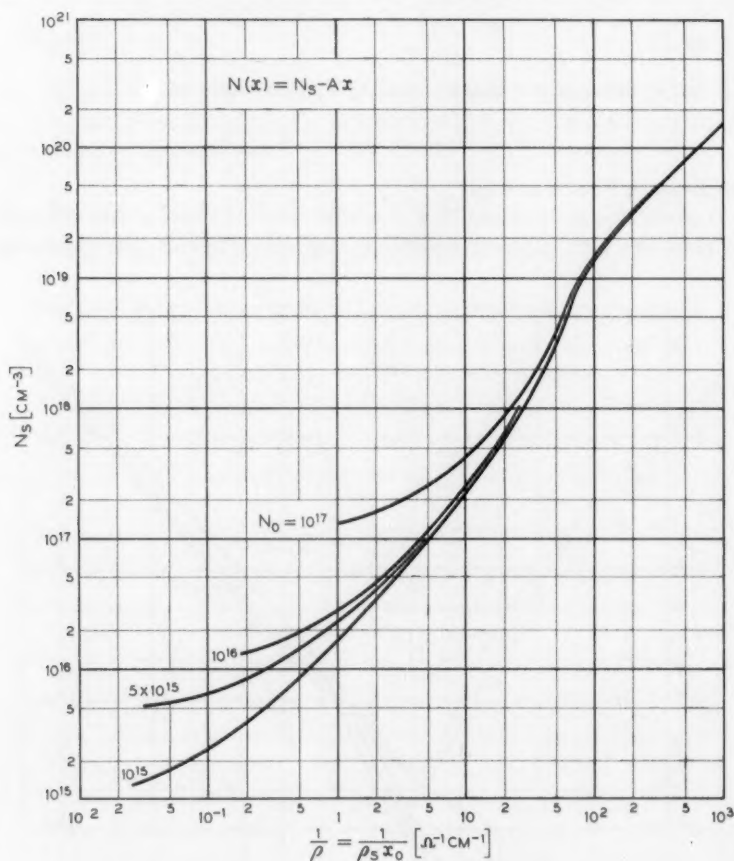
Fig. 9 — Surface concentration N_s versus conductivity for an n-type layer; linear distribution.

energies of all commonly used Group V donor elements differ from each other only slightly. Therefore, (4a) is approximately true for all these elements. The differences in the ionization energies of the Group III acceptor elements are also small, with the exception of indium. Hence, (4b) is good for boron, aluminum and gallium, but is not valid for indium. Deviations of N_s somewhat less than a factor of two may be expected in this case.
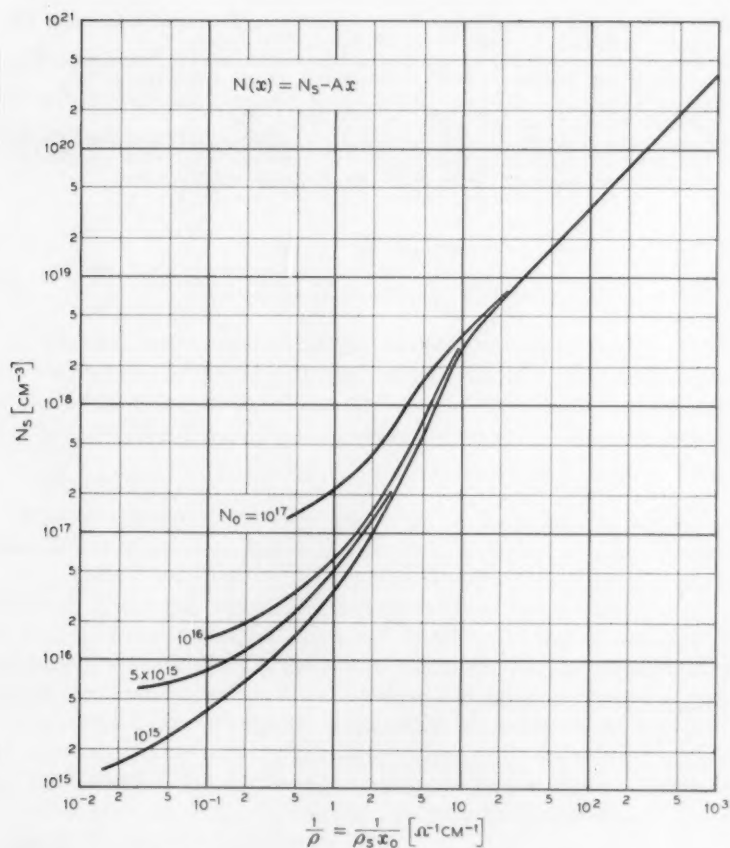
Fig. 10 — Surface concentration N_s versus conductivity for a p-type layer linear distribution.

IV. ACKNOWLEDGMENT

The author wishes to thank Miss M. C. Gray for her assistance in the mathematical computations.

REFERENCES

1. Bond, W. L. and Smits, F. M., The Use of an Interference Microscope for Measurement of Extremely Thin Layers, B.S.T.J., **35**, September, 1956, p. 1209.

2. Fuller, C. S. and Ditzenberger, J. A., Diffusion of Donor and Acceptor Elements in Silicon, J. Appl. Phys., **27**, May, 1956, p. 544.
3. Valdes, L. B., Resistivity Measurements on Germanium for Transistors, Proc. of I.R.E., **42**, February, 1954, p. 420; Smits, F. M., this issue, p. 711.
4. Backenstoss, G., Conductivity Mobilities of Electrons and Holes in Heavily Doped Silicon, Phys. Rev., **108**, Dec. 15, 1957, p. 416.
5. Smits, F. M., Miller, R. C., and Batdorf, R. L., Surface Effects on the Diffusion of Impurities in Semiconductors, in *International Symposium on Semiconductors and Phosphors (1956)*, F. Viweg, Braunschweig, in press.

Measurement of Sheet Resistivities with the Four-Point Probe

By F. M. SMITS

(Manuscript received October 15, 1957)

Correction factors are evaluated for the measurement of sheet resistivities on two-dimensional rectangular and circular samples with the four-point probe. Diffused surface layers can be treated as two-dimensional structures, but the factors are also useful in obtaining body resistivities on thin samples.

I. INTRODUCTION

The "four-point probe" has proven to be a convenient tool for the measurement of resistivities. For a description of the method see a paper by L. Valdes,[1] which gives the functional relationship between the resistivity, ρ, and the voltage and current readings for various geometries. Later, A. Uhlir evaluated functions[2] which give the relationship for additional geometries. All these treatments are concerned with three-dimensional structures infinite in at least one direction.

For diffused layers, a similar relationship is needed for the evaluation of sheet resistivities on various sample shapes. This is a two-dimensional problem which is treated here for various finite sample sizes. With the solution, however, it also is possible to obtain body resistivities on thin slices of the same finite geometries.

II. METHOD

A current source in an infinite sheet gives rise to the logarithmic potential

$$\varphi - \varphi_0 = -\frac{I\rho_s}{2\pi} \ln r,$$

where φ is the potential, I the current, ρ_s the sheet resistivity and r the distance from the current source.

In particular, the potential for a dipole (+ source and − source) becomes

$$\varphi - \varphi_0 = \frac{I\rho_s}{2\pi} \ln \frac{r_1}{r_2}.$$

In the case of a four-point probe on a sheet, the two outside (current) points represent the dipole. Therefore, the potential difference between the two inner points is, for an infinite sheet,

$$\Delta\varphi = V = \frac{I\rho_s}{\pi} \ln 2$$

(only equal point spacing is considered). Thus, the sheet resistivity is obtained as

$$\rho_s = \frac{V}{I} \frac{\pi}{\ln 2} = \frac{V}{I} 4.5324 \cdots .$$

To obtain the sheet resistivity on a finite sample, the method of images can be applied.[3] Only nonconducting boundaries are considered; the boundary (edges) of a diffused layer must be etched or else the layer on the back side of the sample would act as a shunting path.

III. RECTANGULAR SAMPLE

We consider a four-point probe on a rectangular sample with the dimensions a and d. The probe is arranged symmetrically with point spacing s according to Fig. 1. To obtain the voltage between the two center points 1 and 2, an infinite arrangement of dipoles must be considered, as shown in Fig. 2. All contribute to the voltage between points 1 and 2.

F. Ollendorff[4] gives the potential distribution for an infinite number of current sources, arranged in a line and equally spaced. With a coordinate system as in Fig. 3 the potential is

$$\varphi - \varphi_0 = -\frac{I\rho_s}{2\pi} \ln 2 \sqrt{\sin^2 \frac{\pi}{d} x + \sinh^2 \frac{\pi}{d} y}.$$

Fig. 1 — Arrangement of a four-point probe on a rectangular sample.

Fig. 2 — System of images.

With this expression the present problem is reduced to a summation of lines of current sources with alternating sign in only one direction. In the coordinate system for every line of sources the points 1 and 2 have the x-coordinate zero. Thus the expression simplifies to:

$$\varphi - \varphi_0 = -\frac{I\rho_s}{2\pi} \ln \left(e^{\pi y/d} - e^{-\pi y/d} \right).$$

Each line of sources thus contributes to the voltage V the amount

$$\Delta\varphi_n = -\frac{I\rho_s}{2\pi} \ln \left(\frac{e^{\pi(y_n+s)/d} - e^{-\pi(y_n+s)/d}}{e^{\pi y_n/d} - e^{-\pi y_n/d}} \right)$$

where y_n is the distance from point 1 to the center source in line n. The dimension a is involved in the values y_n. The total voltage between the
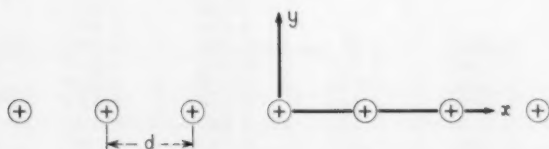
Fig. 3 — Coordinate system for a linear arrangement of current sources.

points 1 and 2 is therefore

$$V = \Sigma\Delta\varphi_n \equiv \rho_s I \frac{1}{C\left(\dfrac{a}{d}; \dfrac{d}{s}\right)}$$

where $C(a/d; d/s)$ is a constant defined by this equation. As shown in the Appendix, the summation can be done easily by expanding the logarithm and summing each term as a geometrical series. In almost all cases, the first term gives four-place accuracy. The sheet resistivity is thus given by
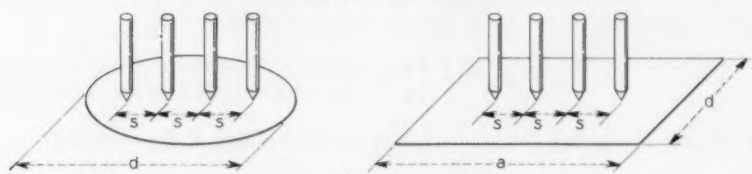
$$\rho_s = \frac{V}{I} C\left(\frac{a}{d}; \frac{d}{s}\right).$$

Table I gives this factor C for various geometries.

For small d/s the quantity $C' = (s/d)C$ is close to unity. In these cases the sheet resistivity can be expressed as:

$$\rho_s = \frac{V}{I}\frac{d}{s}C' \approx \frac{V}{I}\frac{d}{s}.$$

TABLE I — CORRECTION FACTOR C FOR THE MEASUREMENT OF SHEET
RESISTIVITIES WITH THE FOUR-POINT PROBE

$$\rho_s = \frac{V}{I} C$$

d/s	circle diam d/s	$a/d = 1$	$a/d = 2$	$a/d = 3$	$a/d \geq 4$
1.0				0.9988	0.9994
1.25				1.2467	1.2248
1.5			1.4788	1.4893	1.4893
1.75			1.7196	1.7238	1.7238
2.0			1.9454	1.9475	1.9475
2.5			2.3532	2.3541	2.3541
3.0	2.2662	2.4575	2.7000	2.7005	2.7005
4.0	2.9289	3.1137	3.2246	3.2248	3.2248
5.0	3.3625	3.5098	3.5749	3.5750	3.5750
7.5	3.9273	4.0095	4.0361	4.0362	4.0362
10.0	4.1716	4.2209	4.2357	4.2357	4.2357
15.0	4.3646	4.3882	4.3947	4.3947	4.3947
20.0	4.4364	4.4516	4.4553	4.4553	4.4553
40.0	4.5076	4.5120	4.5129	4.5129	4.5129
∞	4.5324	4.5324	4.5324	4.5325	4.5324

TABLE II — CORRECTION FACTOR C' FOR THE MEASUREMENT OF SHEET RESISTIVITIES WITH THE FOUR-POINT PROBE ON NARROW STRUCTURES
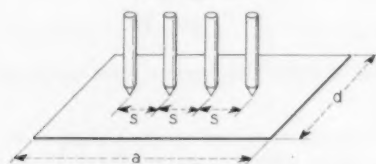
$$\rho_S = \frac{V}{I}\,\frac{d}{s}\,C'$$

d/s	$a/d = 1$	$a/d = 2$	$a/d = 3$	$a/d \geq 4$
1.0			0.9988	0.9994
1.25			0.9973	0.9974
1.5		0.9859	0.9929	0.9929
1.75		0.9826	0.9850	0.9850
2.0		0.9727	0.9737	0.9737
2.5		0.9413	0.9416	0.9416
3.0	0.8192	0.9000	0.9002	0.9002
4.0	0.7784	0.8061	0.8062	0.8062

Table II gives C' for various geometries. The table may be used to determine the error one makes by just approximating with $C' = 1$.

IV. CIRCULAR SAMPLE

In this case, only one image is necessary to fulfill the boundary condition. The image is obtained by "reflecting" the dipole on the circle.[5] For a four-point probe centered in the sample, the voltage becomes:

$$V = \frac{I\rho_s}{\pi}\left[\ln 2 + \ln\left(\frac{\left(\frac{d}{s}\right)^2 + 3}{\left(\frac{d}{s}\right)^2 - 3}\right)\right] = I\rho_s\,\frac{1}{C\left(\frac{d}{s}\right)}\,,$$

where d is the diameter of the circle. The correction factor is also given in Table I.

V. MEASUREMENT OF BODY RESISTIVITIES ON THIN SAMPLES

On a slice of finite thickness w, the four-point probe will introduce voltage gradients perpendicular to the surface. Insofar as these gradients are negligible, the slice can be treated in the same way as an infinitely thin slice and the proper sheet resistivity can be obtained, thus giving the body resistivity ρ with the relation $\rho = \rho_s w$.

As mentioned, for a four-point probe on an infinite sheet the following relation holds:

$$\rho_s = \frac{V}{I}\frac{\pi}{\ln 2}.$$

For an infinite slice of finite thickness w, one can express the resistivity as

$$\rho = \rho_s w = \frac{V}{I}\,w\,\frac{\pi}{\ln 2}\,F\left(\frac{w}{s}\right),$$

where $F(w/s)$ is a correction factor approaching unity as w approaches zero.

From Uhlir's paper[2] $F(w/s)$ can be obtained; it is tabulated in Table III. This table may serve to evaluate the error one makes by treating thin samples as an infinitesimal sheet. For a finite sample, $F(w/s)$ could be used as a first order correction, thus giving the resistivity of a slice as $\rho = \rho_s\,w = (V/I)w\,C\,F$. The exact treatment would require the summation of the potential from an infinite three-dimensional array of dipoles.

TABLE III — MEASUREMENT OF BODY RESISTIVITIES ρ ON THIN SAMPLES OF THICKNESS w
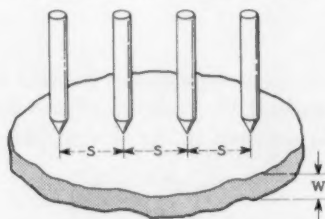
$$\rho = \rho_s \cdot w = \frac{V}{I}\cdot w \cdot \frac{\pi}{\ln 2}\cdot F\left(\frac{w}{s}\right)$$

w/s	$F(w/s)$
0.4	0.9995
0.5	0.9974
0.5555	0.9948
0.6250	0.9898
0.7143	0.9798
0.8333	0.9600
1.0	0.9214
1.1111	0.8907
1.25	0.8490
1.4286	0.7938
1.6666	0.7225
2.0	0.6336

APPENDIX

Summation of $\Delta \varphi_n$

Without loss of generality the following substitutions are made:

$$d \text{ is used for } \frac{d}{s}, \qquad a \text{ for } \frac{a}{s}, \qquad \text{and} \qquad y \text{ for } \frac{y}{s}.$$

From symmetry reasons, it follows that each source in the upper half of the plane (Fig. 2) contributes the same $\Delta \varphi_n$ as does the corresponding source in the lower half. Thus only the sources in the lower half are to be considered and the result must be multiplied by 2.

The terms

$$-\ln (e^{\pi y/d} - e^{-\pi y/d})$$

are first written in the form

$$-\ln e^{\pi y/d}(1 - e^{-2\pi y/d}) = -\pi \frac{y}{d} - \ln (1 - e^{-2\pi y/d}).$$

With this one obtains

$$\frac{2\pi}{|I|\rho_s} \Delta\varphi_n = \pm\left[-\frac{\pi}{d} - \ln (1 - e^{-2\pi(y_n+1)/d}) + \ln (1 - e^{-2\pi y_n/d}) \right].$$

with $+$ standing for a positive source and $-$ for a negative source. Summing the first term gives π/d.

To sum the ln terms the logarithm can be expanded:

$$-\ln (1 - x) = x + \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 + \cdots.$$

With this, each term in n becomes

$$\pm\sum_{m=1}^{\infty} \frac{1}{m} [e^{-2\pi(y_n+1)m/d} - e^{-2\pi y_n m/d}].$$

The y_n can be expressed as

$$y_n = A_i + n2a$$

with four different A_i. With this, each term in m becomes

$$a_m = \sum_{i=1}^{4} \sum_{n=0}^{\infty} \pm \frac{1}{m} (e^{-4\pi am/d})^n e^{-2\pi A_i m/d}[e^{-2\pi m/d} - 1].$$

This is a geometrical series in n. Forming $\sum_{n=0}^{\infty}$ gives

$$a_m = \sum_{i=1}^{4} \pm \frac{1}{m} \frac{(e^{-2\pi m/d} - 1)}{(1 - e^{-4\pi am/d})} e^{-2\pi A_i m/d}.$$

For reasons of convergence, $y = 1$ is treated separately. The A_i thus become

for $+$ sources:

$$A_2 = a + 1$$
$$A_4 = 2a - 2$$

for $-$ sources:

$$A_1 = 2a + 1$$
$$A_3 = a - 2.$$

Forming $\sum_{i=1}^{4} (-1)^i e^{-2\pi A im/d}$ and inserting into the expression for a_m gives

$$a_m = \frac{1}{m} e^{-2\pi(a-2)m/d} \frac{(1 - e^{-6\pi m/d})(1 - e^{-2\pi m/d})}{(1 + e^{-2\pi am/d})}.$$

The total voltage between points 1 and 2 is therefore

$$V = I\rho_s \frac{1}{\pi} \left[\frac{\pi}{d} + \ln (1 - e^{-4\pi/d}) - \ln (1 - e^{-2\pi/d}) + \sum_{m=1}^{\infty} a_m \right].$$

In all but three cases the first term in m gave four-place accuracy.

REFERENCES

1. Valdes, L., Resistivity Measurements on Germanium for Transistors, Proc. I.R.E., **42**, Feb., 1954, p. 420.
2. Uhlir, A., Jr., The Potentials of Infinite Systems of Sources and Numerical Solutions of Problems in Semiconductor Engineering, B.S.T.J., **34**, Jan., 1955, p. 105.
3. Thomson, Sir Willam, Reprint of *Papers on Electrostatics and Magnetism*, Macmillan, London, 1884.
4. Ollendorff, F., *Potentialfelder der Elektrotechnik*, Springer, Berlin, 1932.
5. See for example: *Encyclopädie der Mathematischen Wissenschaften*, **5**, 2. Teil, Teubner B. G., Leipzig, 1904–1922, p. 422.

Piezoelectric and Dielectric Characteristics of Single-Crystal Barium Titanate Plates

By A. H. MEITZLER and H. L. STADLER

(Manuscript received October 18, 1957)

The dielectric constant of c-domain barium titanate single-crystal plates has been measured as a function of their average polarization, using ac voltages much smaller than the coercive voltage. As has been found previously, the dielectric constant at frequencies below thickness fundamental resonance has a minimum at zero average polarization. However, at frequencies well above thickness resonance the dielectric constant has a maximum at zero polarization. The variation is about 25 per cent from minimum to maximum in both frequency ranges. In both frequency ranges the conductance is highest at zero polarization.

The piezoelectric output voltage of a $BaTiO_3$ single crystal subjected to a recurrent strain pulse of constant amplitude has been found to be approximately, but not exactly, proportional to its average polarization. There appears, in fact, to be a distinct hysteresis in the dependence.

I. INTRODUCTION

The physical properties of $BaTiO_3$ have been investigated extensively in the last few years, and the ferroelectric characteristics of single crystals have found several applications in electrical circuit devices. On the other hand, applications of the piezoelectric characteristics of $BaTiO_3$ have largely been limited to its ceramic form. The experiments reported in this paper were initially undertaken to investigate the possible usefulness of the piezoelectric characteristics of $BaTiO_3$ single crystals in practical devices.

In the work reported in this paper, dielectric and piezoelectric characteristics of $BaTiO_3$ in single crystals were investigated by means of two types of electrical measurements. A crystal plate with electrodes is essentially a two-terminal circuit element. The admittance of crystals as passive elements has been measured as a function of average remanent

polarization above, below and in the vicinity of the fundamental resonant frequency in the thickness-extensional mode. When such a crystal is excited by ultrasonic radiation it becomes a voltage source. The output voltage of an ultrasonically excited crystal has been observed as a function of average remanent polarization when the ultrasonic excitation was at frequencies in the vicinity of the fundamental resonant frequency in the thickness-extensional mode.

Certain features of the piezoelectric behavior of $BaTiO_3$ single crystals have been measured before by Bond, Mason and McSkimin;[1] by Caspari and Merz;[2] by Drougard and Young[3] and by Huibregtse, Drougard and Young.[4] Measurements having particular bearing on the work reported in this paper were made by Drougard and Young, who measured the low-frequency, small-signal capacitance of $BaTiO_3$ crystals as a function of their average polarization. This type of measurement receives further elaboration in the present paper.

When comparing reported results of measurements of dielectric and other material constants, one should bear in mind that the quality of $BaTiO_3$ single crystals may vary considerably. The dielectric constant is a rather good index of the fraction of a-domains in these crystals because it is much higher for a-domains than for c-domains. The low-frequency dielectric constant $(K_{33}{}^T)$ was earlier reported to have values ranging from 185 to approximately 3000. These values should be compared with later values found by Merz[5] of 160 and our own values, which are in this same range.

For applications of the type we wished to consider, plate-shaped, c-domain crystals were most suitable. Plate-shaped crystals were grown by the method of Remeika[6] and were rendered entirely c-domain by application of an electric field perpendicular to the plate. Coin silver electrodes $\frac{3}{16}$ inch in diameter and about 1000 Å thick were evaporated onto the broad sides of the plate.

Some crystals were bonded with Araldite to fused silica rods before being measured. The operations of electroding and bonding, where they occurred, could possibly have introduced a-domains, but these could not be detected by polarized light because both the electrodes and the silvered surfaces of the fused silica were opaque. However, the low values of dielectric constant lead us to believe that the plates were very nearly pure c-domain during our measurements.

In the course of the experiments to be reported, measurements were made on several dozen $BaTiO_3$ single crystals having different thicknesses and electrode configurations. From the reader's point of view it would be simplest if all the different types of measurements reported were reported

on the same crystal. This was not practical, but the number of samples
has been limited to three crystals, which have been designated as crys-
tals "A", "B" and "C". For these crystals, certain quantities were
measured as functions of polarization, and a complete set of numerical
results for dielectric and piezoelectric quantities is presented in Table I
(see page 731). All the crystals observed, including the three selected
crystals, exhibited variations in behavior, due perhaps to undetected
differences in processing. For measurements where the results varied
considerably from crystal to crystal, the variations will be specifically
mentioned.

II. EXPERIMENTAL METHODS AND RESULTS

2.1 *Method of Varying Average Polarization*

The average polarization of a BaTiO$_3$ crystal was changed by means
of the circuit shown in Fig. 1(a).[7] The value of C_1 was chosen so that a
small amount of charge, C_1V_B , could be moved through the crystal each
time the switch S_1 was moved from the shorting position to the position
connecting the capacitor in series with the crystal; C_1 was chosen so
that C_1V_B was approximately $\frac{1}{35} Q_r$, where Q_r is the remanent charge.
For the crystals used in these experiments, Q_r was approximately 5
microcoulombs, C_1 was 0.002 microfarads and V_B was 67 volts. The
changes in the crystal's state of polarization produced by the circuit of
Fig. 1(a) may be interpreted in terms of the diagram shown in Fig. 1(b).
The crystal may be put in the state of polarization indicated by the point
p_1 on the loop by applying a voltage which is greater than the coercive
voltage. When the crystal is shorted, it changes to position p_2 on the
hysteresis loop. When an amount of charge equal to C_1V_B has passed
through the series circuit all the voltage appears across C_1 and the crystal
is now in position p_3 . The charge accumulated on C_1 can be shorted out
by S_1 and the switching process can be repeated until the polarization
saturates in the opposite direction. Reversing the applied voltage will
return the crystal, a step at a time, to position p_2 . For some observa-
tions, the crystal was also switched slowly and continuously by the use
of a large resistance (of the order of 10^{10} ohms) in place of C_1 . In either
case, the charge moved through the crystal was measured by the elec-
trometer E across the low-leakage capacitor C_2 .

2.2 *Admittance Measurements*

In general, the admittance of a dielectric material between two elec-
trodes is a complicated function of frequency, depending upon the pres-

ence of various types of physical phenomena such as piezoelectricity, ferroelectricity, relaxation effects or electrostriction. In the case of a given sample of material, measurements of admittance over a wide range in frequency are used to establish which physical phenomena are present. In addition to having a dependence upon frequency, values of admittance are frequently dependent upon the curent level at which the measurements were made. This is true for $BaTiO_3$ single crystals, which have the added complication that the admittance at a particular frequency de-

(a)

BASIC CIRCUIT USED TO CHANGE STATE OF POLARIZATION
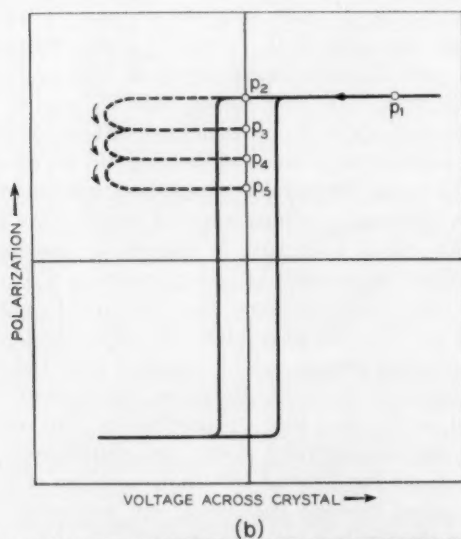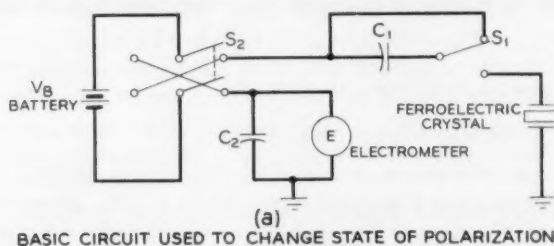
(b)
AN IDEALIZED HYSTERESIS LOOP OF A FERROELECTRIC CRYSTAL AND THE PATH FOLLOWED IN CHANGING THE STATE OF POLARIZATION

Fig. 1 — (a) Circuit for switching small amounts of charge through a ferroelectric crystal; (b) diagram of crystal's hysteresis loop showing path traced by circuit of (a).

pends upon the state of polarization. It is possible (and desirable for many applications) to switch a crystal to an arbitrary polarization by means of the type of circuit shown in Fig. 1(a). Admittance may then be measured as a function of frequency for that particular value of average polarization. Provided that the measuring circuit does not place across the crystal a voltage large enough to move domains, the crystal will stay at the set value of average polarization.

All the admittance measurements reported in this paper were made with measuring voltages across the crystal small enough so that the polarization of the crystal was left unchanged by a measurement. With this sort of technique we could, therefore, measure admittance at a given frequency as a function of polarization or measure admittance at a given polarization as a function of frequency.

2.2.1 *Admittance Measurements in the Vicinity of the Fundamental Resonant Frequency of the Thickness-Extensional Mode.* The admittance of a dielectric material having piezoelectric characteristics varies greatly with frequency when it is near resonance, compared with the frequency dependence of the admittance far removed from a resonant frequency. For this reason, admittance measurements fall rather naturally into two major types: (1) those near resonance and (2) those far removed from resonance.

In order to define several terms of which we will make frequent use, we show in Fig. 2 the admittance characteristics of a simple piezoelectric resonator having a frequency dependence of admittance that closely approximates the behavior of our $BaTiO_3$ crystals at maximum polarization. In this figure the admittance of the resonator at any frequency, f, is given in terms of its capacitance, $C_E(f)$, and its shunt resistance $R_E(f)$ according to

$$Y(f) = \frac{1}{R_E(f)} + j\omega C_E(f). \tag{1}$$

The series resonant frequency, f_s, is defined as that frequency at which $R_E(f)$ is a minimum. Two other characteristic frequencies are f_r, the resonant frequency, and f_a, the antiresonant frequency. At f_r, $C_E(f)$ equals zero and dC_E/df is negative; at f_a, $C_E(f)$ is again zero and dC_E/df is positive.

The resonant and antiresonant frequencies in the thickness-extensional mode of the $BaTiO_3$ crystals described above were measured by an admittance bridge technique described elsewhere.[3] After the average polarization was set by the circuit of Fig. 1(a), the crystal was removed to the admittance bridge for measurement. The value of the polarization,
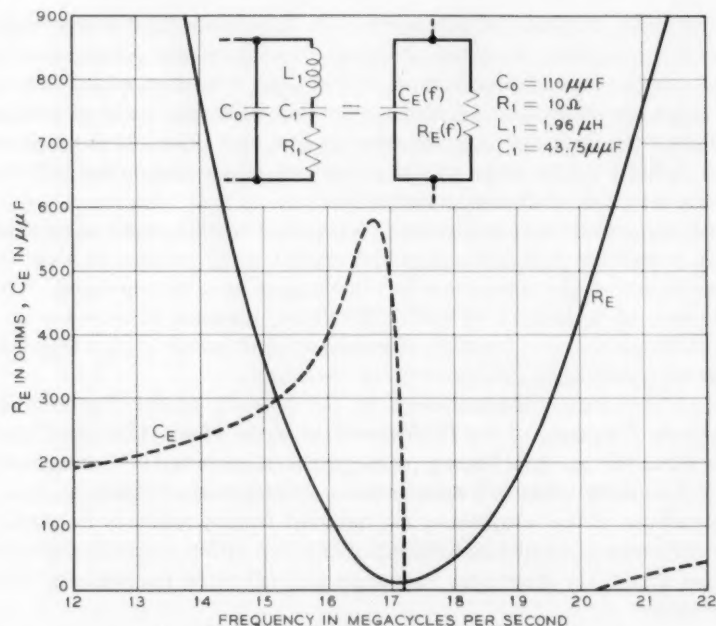
Fig. 2 — Admittance characteristics of a simple piezoelectric crystal near its resonant frequency.

measured by the electrometer, was found to be unchanged after the admittance measurement.

The admittance of an unbonded crystal was measured as a function of frequency, both in a state of maximum polarization and in a state of nearly zero polarization. The capacitance and shunt resistance are plotted against frequency in Fig. 3 for maximum polarization and in Fig. 4 for nearly zero average polarization. The many small erratic variations in the curves of Fig. 3, as compared with those of Fig. 2, are attributed to the presence of multiple resonances possibly due to the uneven surfaces* of the crystal plate. Fig. 5 shows a plot of resonant and antiresonant frequencies of the same crystal as a function of polarization.

Measurements exactly the same as those made on unbonded crystals were made on a similar crystal which had been bonded with Araldite

---

* The surfaces of the crystal plates were not lapped. They were etched to a convenient thickness by concentrated phosphoric acid and given a final wash in dilute hydrochloric acid.
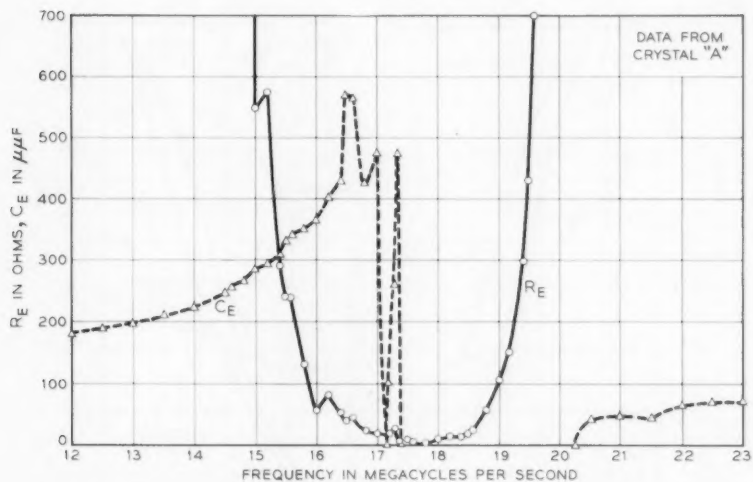
Fig. 3 — Admittance of an unbonded BaTiO₃ single crystal at maximum polarization near its resonant frequency.
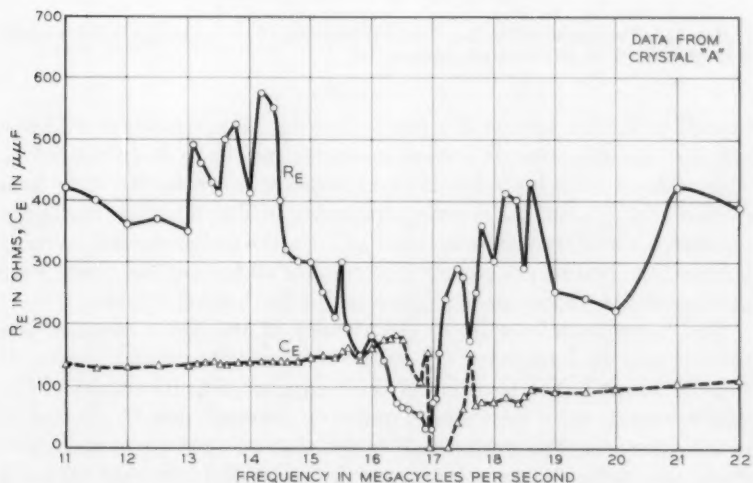
Fig. 4 — Admittance of an unbonded BaTiO₃ single crystal near zero average polarization in the vicinity of its resonant frequency.
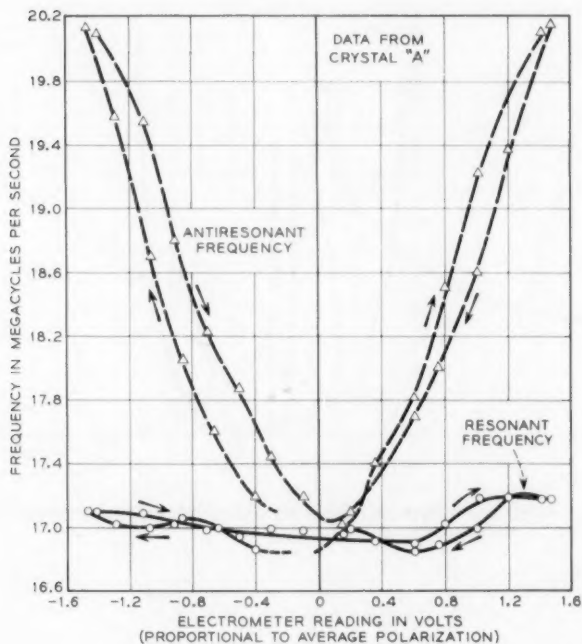
Fig. 5 — Resonant and antiresonant frequencies of an unbonded BaTiO$_3$ single crystal plotted vs. its average polarization.

cement to the flat surface of a fused silica rod. This bonded crystal was also used for piezoelectric voltage measurements to be described below. The results of the measurements near resonance on crystal "C" have been plotted as $C_E(f)$ and $R_E(f)$ versus frequency in Fig. 6 for the maximum remanent polarization state and in Fig. 7 for the state near zero average polarization. Since C_E is greater than zero for all frequencies, there were no resonant or antiresonant frequencies for the bonded crystal.

The experimental results in the vicinity of the series resonant frequency may be interpreted by means of equivalent circuits shown on Figs. 2, 6 and 7. The elements of the equivalent circuits consist of C_0, which depends upon the clamped dielectric constant, and R_1, C_1 and L_1 in the series arm. The quantities L_1 and C_1 depend upon the piezoelectric, elastic and dielectric characteristics of the material. In an ideal unloaded piezoelectric resonator, R_1 is zero. A non-zero value of R_1 for a real unbonded crystal is the result of energy being dissipated by the mount-
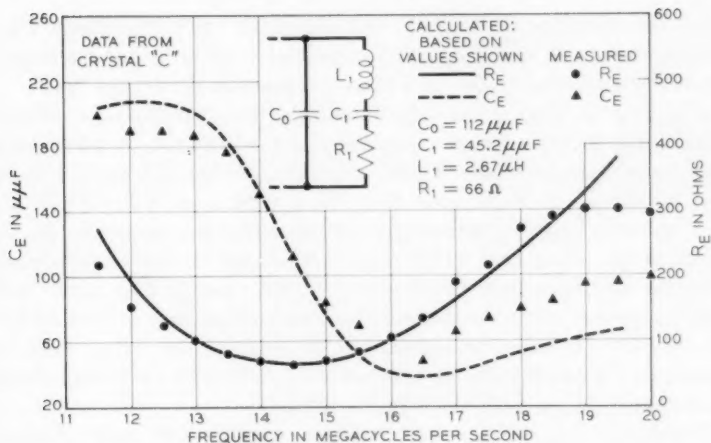
Fig. 6 — Admittance of a bonded BaTiO₃ single crystal at maximum polarization near its resonant frequency.
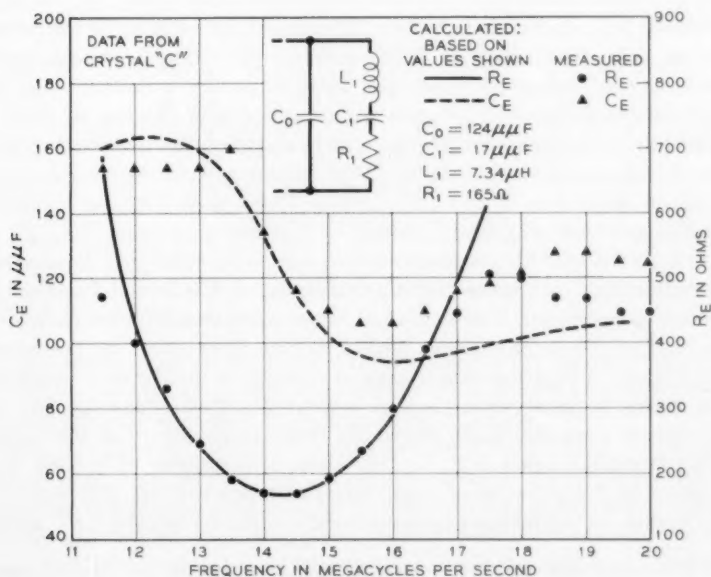
Fig. 7 — Admittance of a bonded BaTiO₃ single crystal near zero average polarization in the vicinity of its resonant frequency.

ing of the crystal or by anelastic mechanisms in the crystal itself. For a bonded crystal, R_1 includes the effects of the loading by the medium to which the crystal is bonded as well as the losses in the crystal itself.

The data in Figs. 3 through 7 indicate that variations in average polarization produce major variations in the admittance of a crystal at frequencies near resonance. These variations are caused in part by variations in the effective clamped dielectric constant, as indicated by C_0. The relatively large variations are due primarily to changes in the elements in the series arm which depend upon the effective piezoelectric constant and upon loss mechanisms in the crystal. The capacitance ratio* increases when the polarization approaches zero, indicating that the effective piezoelectric constant is decreasing. The value of R_1 increases as the polarization approaches zero, indicating that more energy is dissipated during each cycle by the crystal.

The variation in effective piezoelectric constant is most clearly revealed by the curves of Fig. 5, which show the variation in resonant and antiresonant frequencies as a function of polarization. When the polarization approaches zero, the separation between the two frequencies decreases. Because the square of the effective piezoelectric coefficient, d_{33},† is proportional to $f_a - f_r$ or Δf, these data indicate a decrease in d_{33} as the polarization approaches zero. It is possible for the effective piezoelectric coefficient to become so small that there is neither a resonant nor an antiresonant frequency; i.e., the variation in C_E is reduced so that it never passes through zero. Fig. 5, in the case when the polarization was varied from positive to negative values, shows gaps in the curves for f_r and f_a in the region of zero polarization. These gaps were the result of the disappearance of f_r and f_a caused by the near-zero value of d_{33}.

2.2.2 *Admittance Measurements at Frequencies Removed from Resonance.* The admittance of crystal "A" was measured at 100 kc as a function of average polarization. The results of these measurements are shown in Fig. 8. The 100-kc plot agrees with a similar observation by Drougard and Young[3] in that the capacitance of crystals in unsaturated polarization states is lower than that in saturated states. Measurements of admittance were also made above the resonant frequency in the thickness-extensional mode and, in this case, a surprising difference was observed in the variation of capacitance with polarization. Fig. 9 shows the results for admittance measurements made on crystal "B" at 26

---

* The capacitance ratio, r, is defined in terms of elements in the equivalent circuit as C_0/C_1. The elements L_1 and C_1 are interrelated by the requirement that $\omega_s{}^2 = 1/L_1 C_1$.

† The symbols used for all elastic and electric quantities will be those given by the IRE Standards on Piezoelectric Crystals, 1949, unless otherwise noted.
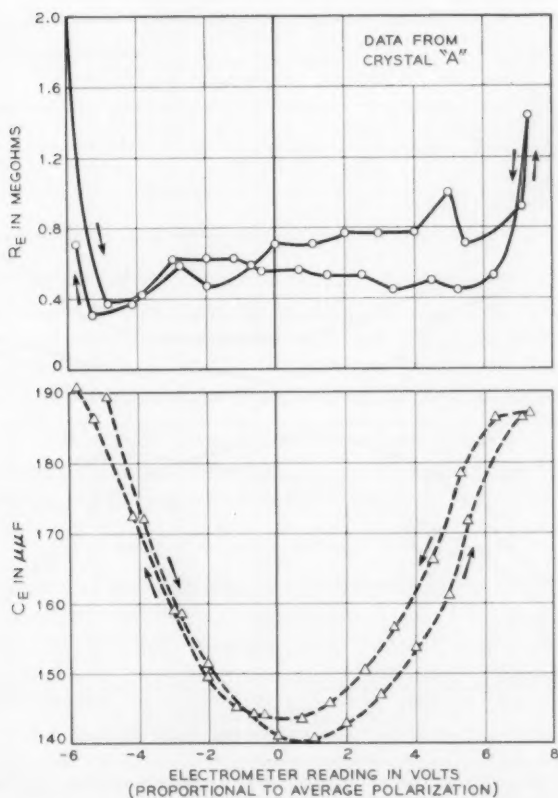
Fig. 8 — Admittance of an unbonded BaTiO₃ single crystal at 100 kc vs. its average polarization.

me, which is approximately twice the fundamental resonant frequency of this crystal. It can be seen that the dielectric constant for zero polarization is higher than that for maximum polarization.

The data shown in Figs. 8 and 9 are representative of the results obtained on our better crystals for the variation in the small signal dielectric constant as a function of polarization and of frequency. These results will now be summarized and given a qualitative interpretation.

At low frequencies, the capacitance, and hence the dielectric constant, is roughly a parabolic function of polarization, being maximum at the extremes in polarization and having at zero polarization a minimum which is down 10 to 25 per cent from the maximum value. At higher
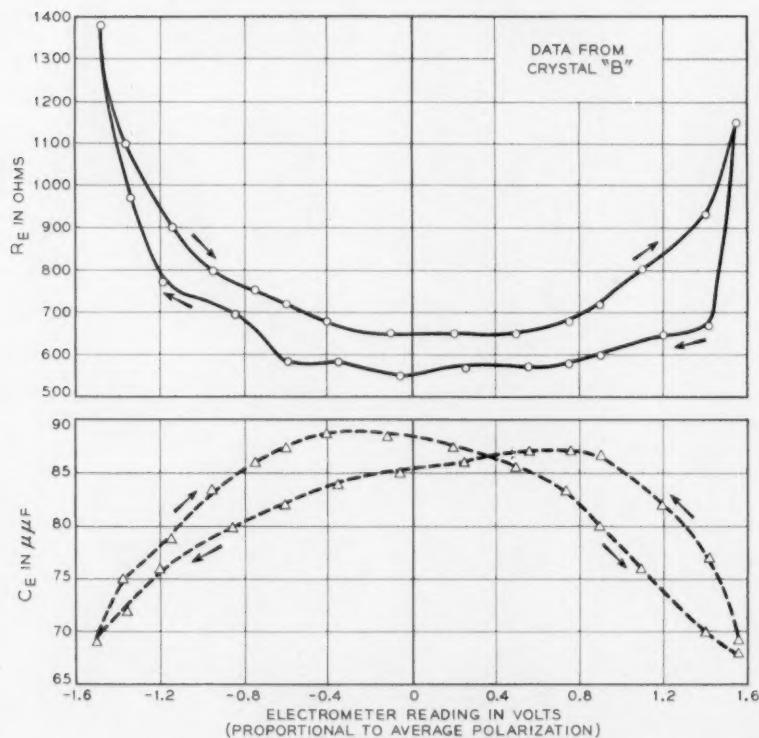
Fig. 9 — Admittance of an unbonded BaTiO₃ single crystal at 26 mc vs. its average polarization.

frequencies, the dielectric constant is again approximately a parabolic function of polarization, but in this case the parabola is turned downward; i.e., the maximum value of the dielectric constant occurs at zero polarization and the minimum values occur at the extremes in polarization. On the high-frequency side of resonance, the minimum values have been observed to be 20 to 30 per cent below the maximum value at zero polarization. Values for dielectric constants calculated for three BaTiO₃ single crystals are given below. These values have not been corrected for the fringing polarization which was present, since the samples were not fully electroded. The effect of such a correction would be to lower the values of the dielectric constants; consequently, the values given in Table I should be regarded as upper limits on the corrected values.

TABLE I

Crystal	f_r (mc)	$K_{LF,MP}$	$K_{LF,ZP}$	$K_{HF,MP}$	$K_{HF,ZP}$
A	17.18	222	165	94	129
B	12.85	132	111	57	79
C	14.35	169	145	98	122

The symbols used in Table I are defined as follows:

$K_{LF,MP}$ = low-frequency dielectric constant at maximum polarization,

$K_{LF,ZP}$ = low-frequency dielectric constant at zero polarization,

$K_{HF,MP}$ = high-frequency dielectric constant at maximum polarization,

$K_{HF,ZP}$ = high-frequency dielectric constant at zero polarization.

There is considerable variation in the values of the dielectric constants, and we attribute this to variations in the structures of the individual crystals, but in each case it is true that

$$K_{LF,MP} > K_{LF,ZP} > K_{HF,ZP} > K_{HF,MP} . \qquad (2)$$

The polarization dependent behavior observed at low frequencies has received an interpretation by Drougard and Young.[3] They suggest that the decrease in dielectric constant with decreasing polarization is a consequence of the arrangement of many antiparallel domains in the crystal. The extensional strains developed in these domains under the influence of an applied field oppose each other, resulting in a reduction in the magnitude of strain and consequent reduction in the total energy stored in elastic deformation of the crystal. The crystal is taken as "thickness-clamped" at zero polarization and at frequencies below resonance. Drougard and Young did not consider variations in the dielectric constant at high frequencies with variations in polarization.

In order to provide a qualitative explanation for the difference between $K_{LF,ZP}$ and $K_{HF,MP}$, we suggest that the interpretation of Drougard and Young be supplemented by the inclusion of shear strains at the 180° domain walls. When a low-frequency signal is applied to a crystal in the zero polarization state, the total energy stored in strain is less than at maximum polarization. However, the crystal cannot be regarded as completely clamped, because an appreciable amount of energy is being stored both in shear strains (S_4 and S_5) as well as in extensional strain (S_3). The deformation of a crystal at zero polarization by an applied field is pictured in Fig. 10. When a crystal is measured at high frequencies and maximum polarization, the energy stored in strain is least, because the

amplitude of extensional strain is low and no shear strains are present. While the pictured mechanism may account for the difference between $K_{LF,ZP}$ and $K_{HF,MP}$, it does not explain why $K_{HF,ZP}$ was observed to be greater than $K_{HF,MP}$.

In the present work, the fact that both $K_{HF,ZP}$ and $K_{HF,MP}$ are considerably lower than $K_{LF,ZP}$ shows that a crystal is not very well clamped at low frequencies and zero polarization. For a given crystal, $K_{HF,MP}$ appears to be the best estimate one can make of the clamped dielectric constant, and even it is subject to some doubt. For instance, since the capacitance appears to be still rising with frequency in this range, as can be seen from Fig. 3, the clamped dielectric constant may be higher than $K_{HF,MP}$. On the other hand, since transducers mounted on ultrasonic delay lines can generate mechanical vibrations at these same frequencies, the crystal cannot be completely clamped. Consequently, $K_{HF,MP}$ may be higher than the truly clamped dielectric constant. Finally, since the values of the four dielectric constants varied so much from one crystal to another, it is clear that the crystals used do not provide a good basis for determining the numerical values of the constants.
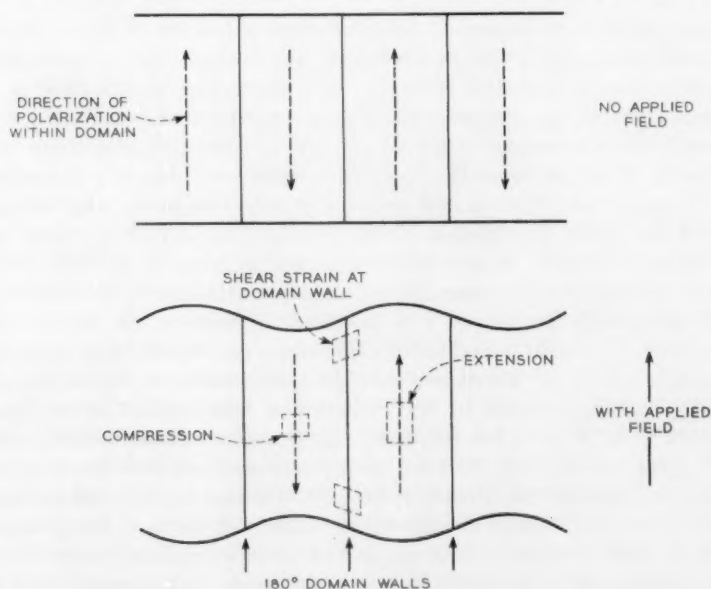
Fig. 10 — Diagram of the deformations produced in a BaTiO$_3$ single crystal near zero average polarization by a small applied field.

Before leaving the subject of admittance measurements, we should observe two other noteworthy features of the data. First, the crystals dissipate more energy per cycle in the zero polarization state than in the state of maximum polarization. This statement is based on the observation that, away from resonance, R_E is always lower for zero polarization than for maximum polarization. (See, for example, Figs. 8 and 9.) Bond, Mason and McSkimin[1] have reported that the material is mechanically more lossy in the zero polarization state. They attributed this extra loss to a microhysteresis effect.

The second observation to be made regards the multiple-valuedness of many of the quantities measured as functions of polarization. The data of Figs. 5, 8 and 9 indicate that the components of the admittance measured as functions of polarization have different values well outside the limits of experimental error, although the curve for any quantity retains the same general shape. This scatter probably indicates that the measured values of these quantities are sensitive to the strain condition of the crystal which, in turn, is dependent upon the exact shape and position of the microscopic domains.
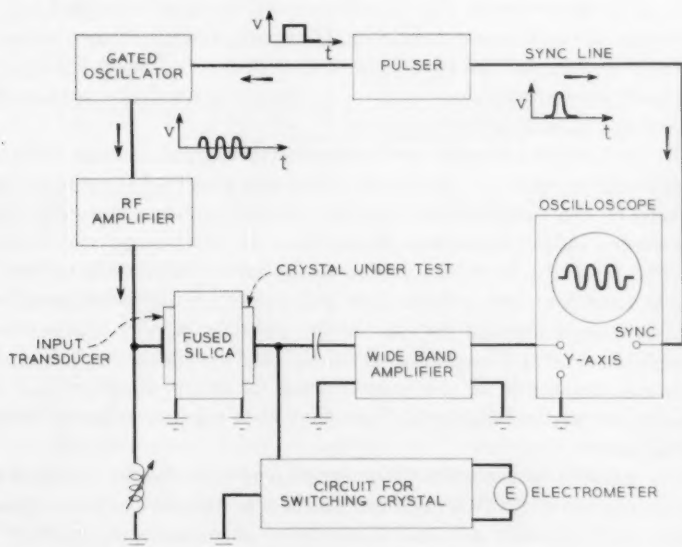
Fig. 11 — Block diagram of apparatus for measuring piezoelectric output voltage of a BaTiO₃ single crystal as a function of its average polarization.

III. MEASUREMENTS OF PIEZOELECTRIC VOLTAGE GENERATED BY ULTRA-SONIC EXCITATION

Two BaTiO$_3$ single crystals were bonded with Araldite to the ends of a fused silica rod, forming an ultrasonic delay line. One of the crystals, crystal "C", was used as the output transducer and its piezoelectric output voltage was measured as a function of its average polarization.

The essentials of the experimental arrangement are shown in Fig. 11. A recurrent RF pulse of constant amplitude and frequency excited the input crystal transducer. The elastic dilatational wave traveled the length of the line and excited a piezoelectric response in the output crystal transducer, which was observed on the oscilloscope. The average polarization of the output transducer was set and measured by the circuit of Fig. 1(a).

Fig. 12 shows the dependence of piezoelectric output voltage on average polarization for crystal "C". Other crystals were measured and the shapes of the resulting curves are similar. The shape of the curve for crystal "C" was found to be very insensitive to changes in the frequency or amplitude of the applied mechanical wave. The decreased slope at the ends of the curve in Fig. 12 is due to the increase in capacitance of the crystal at maximum polarization. However, the capacitance variation does not account for the hysteresis of the curve. Almost all the crystals measured showed this same type of hysteresis dependence of piezoelectric voltage on average polarization.

The vertical jumps near zero piezoelectric output voltage were not found for all crystals. On one crystal there was a vertical jump in traversing zero in one direction but not in the opposite direction. When the piezoelectric output signal was displayed on an oscilloscope and observed very carefully, the waveform of the piezoelectric pulse could be seen to vary in time near zero polarization and a distinct minimum amplitude could be moved through the time of the pulse by passing charge slowly through the crystal. The data used for plotting the curve were always the maximum amplitude of the pulse during its entire duration, and the meaning of vertical jumps is just that this maximum never passed through zero.

It is possible to interpret these experimental results in terms of the domain picture of BaTiO$_3$ crystals. The simple domain picture of polarization would predict a linear dependence of piezoelectric voltage on average polarization. This can be seen as follows. If the crystal consists of many domains polarized in either of the two directions perpendicular to the electrodes on the large surface of the plate, and if the polarization
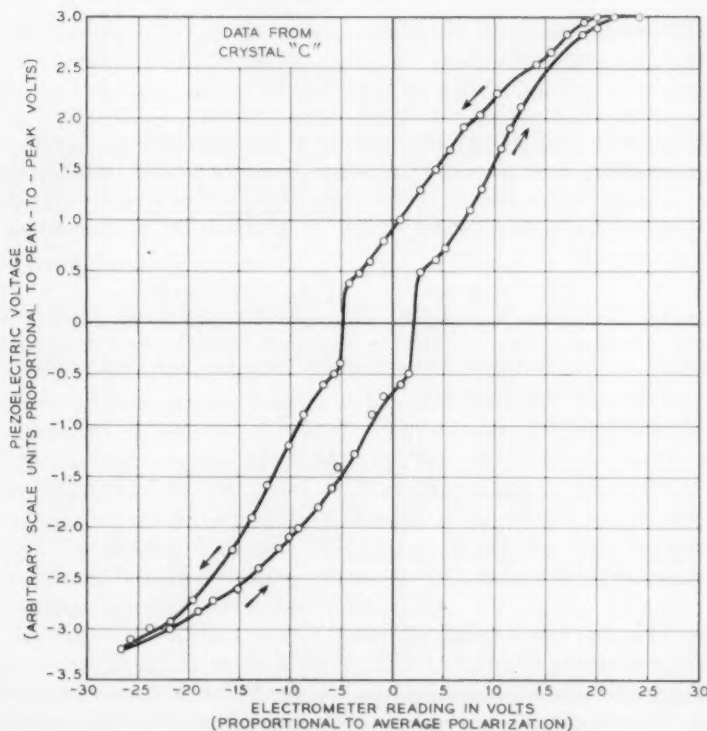
Fig. 12 — Piezoelectric output voltage vs. average polarization of a BaTiO₃ single crystal subjected to constant amplitude strain waves.

per unit volume within each domain has the same value, p, then the average polarization is

$$P = \frac{1}{V_c} \int_{V_c} pndv = p \frac{V_+ - V_-}{V_c}, \qquad (3)$$

where V_c is the volume of the crystal between the electrodes, V_+ and V_- are the volumes of domains polarized positively and negatively respectively, n is a unit vector perpendicular to the electrode and $p = |p|$. The average polarization is indicated in the experiment as

$$P = \frac{C_2 V_{\text{electrometer}}}{A}, \qquad (4)$$

where A is the electrode area.

Thus we have

$$V_{elec} = \frac{pA}{C_2} \frac{V_+ - V_-}{V_e}. \tag{5}$$

We assume that p is a function only of the temperature, T, and the average strain, $S = (L - L_0)/L_0$, where L and L_0 are the instantaneous and average thicknesses of the crystal. Thus, the piezoelectric voltage is assumed to be due to the change in polarization accompanying a strain, as

$$V_{piezo} = \frac{A}{C_x} \frac{\partial P}{\partial S} \Delta S = \frac{A}{C_x} \frac{\partial p}{\partial S} \Delta S \frac{V_+ - V_-}{V_e}, \tag{6}$$

where C_x is the capacitance of the crystal plus the input capacitance of the voltage measuring circuit.

Thus one obtains

$$\frac{V_{piezo}}{V_{elec}} = \frac{\dfrac{A}{C_x} \dfrac{\partial p}{\partial S} \Delta S \dfrac{V_+ - V_-}{V_e}}{\dfrac{pA}{C_2} \dfrac{V_+ - V_-}{V_e}} \tag{7}$$

$$= \frac{C_2 \dfrac{\partial p}{\partial S} \Delta S}{p C_x}.$$

For constant strain and temperature, all the factors in this last relation, with the exception of C_x, are constant, independent of average polarization; C_x is a slowly varying function of average polarization. Thus, we would expect the piezoelectric voltage to be proportional to the average polarization. As Fig. 12 shows, this is not entirely true.

According to the simple picture just presented, we would expect the piezoelectric voltage for maximum positive average polarization to equal that for maximum negative polarization. This was true for some crystals, including the one measured for Fig. 12, but was not true for many others. A plausible cause of this inequality is that some domains are stuck, and are not switched even when the crystal is saturated. That is, V_+, for example, might never be zero, thus reducing the maximum piezoelectric voltage for negative polarization. These stuck domains would also reduce the switching charge of the crystal from its maximum value. In fact, it was found that all crystals whose positive and negative piezoelectric voltages were very unequal did not switch as much charge as they had switched shortly after plating.

As a final observation, it should be noted that, although the simple

dipole theory leads one to expect a linear relationship between piezo-
electric voltage and polarization, what is usually observed is the open-
loop type of curve shown in Fig. 12. Upon repeating the polarization
cycle in measuring a given crystal, one finds that the individual points
are duplicated almost exactly. In this case there seems to be a true
hysteresis, as contrasted to the results obtained for admittance meas-
urements as functions of polarization.

IV. CONCLUSIONS

The dielectric and piezoelectric constants of c-domain $BaTiO_3$ single-
crystal plates are sensitive functions of their average polarization when
determined from capacitance measurements using ac voltages much
smaller than the coercive voltage. It has been reported[3] that the dielec-
tric constant at frequencies below thickness resonance has a minimum
at zero polarization, and our measurements confirm this result. On the
other hand, at frequencies above the fundamental thickness-resonant
frequency, the dielectric constant has a distinct maximum at zero polari-
zation. The dielectric constant varies about 25 per cent from minimum
to maximum in both frequency ranges. In both ranges, the conductance
is maximum at zero polarization.

The dielectric constant of $BaTiO_3$ at high frequencies and maximum
polarization is significantly lower than the dielectric constant at zero
polarization and low frequencies. This difference indicates that, at low
frequencies and zero polarization, the crystal cannot be regarded as com-
pletely clamped, and for this reason the dielectric constant at high
frequencies and maximum polarization represents a better value to use
for the effective dielectric constant at constant strain.

The piezoelectric output voltage of a $BaTiO_3$ single crystal subjected
to constant-amplitude recurrent strain pulses has been measured and
found to be approximately proportional to its average polarization. How-
ever, departures from linearity were observed. When piezoelectric output
voltage was plotted as the ordinate and polarization as the abscissa, a
hysteresis dependence was observed. In addition, it was found that the
curve of piezoelectric output voltage showed vertical jumps in the
vicinity of zero polarization.

The measurements presented in this paper indicate that plate-shaped
$BaTiO_3$ single crystals have a significantly lower dielectric constant and
a smaller capacitance ratio when used as resonators in the thickness-
extensional mode than do similar resonators made from $BaTiO_3$ ceramic.
These characteristics of the single crystal can be used to advantage in

certain applications such as electromechanical transducers for operation at high frequencies.

REFERENCES

1. Bond, W. L., Mason, W. P. and McSkimin, H. J., Elastic and Electromechanical Coupling Coefficients of Single Crystal Barium Titanate, Phys. Rev., **82,** May 1, 1951, pp. 442–443.
2. Caspari, M. E. and Merz, W. J., The Electromechanical Behavior of BaTiO$_3$ Single Domain Crystals, Phys. Rev., **80,** December 15, 1950, pp. 1082–1089.
3. Drougard, M. E. and Young, D. R., Domain Clamping Effect in Barium Titanate Single Crystals, Phys. Rev., **94,** June 15, 1954, pp. 1561–1564.
4. Huibregtse, E. J., Drougard, M. E. and Young, D. R., Electromechanical Behavior of Single Crystals of Barium Titanate, *Annual Report, 1954, Conference on Electrical Insulation* (National Research Council, March 15, 1955), p. 16.
5. Merz, W. J., Double Hysteresis Loop of BaTiO$_3$ at the Curie Point, Phys. Rev., **91,** August 1, 1953, pp. 513–517.
6. Remeika, J. P., A Method for Growing Barium Titanate Single Crystals, J. Am. Chem. Soc., **76,** February 5, 1954, pp. 940–941.
7. A very similar method has been reported by Cameron, D. P., Domain Orientation in Barium Titanate Single Crystals, IBM J. of Research and Development, **1,** January, 1957, pp. 2–7.
8. Meitzler, A. H., A Procedure for Determining the Equivalent Circuit Elements Representing Ceramic Transducers used in Delay Lines, *Proceedings 1957 Electronic Components Symposium*, (May, 1957, Chicago,), pp. 210–219.

Organic Deposits on Precious Metal Contacts

By H. W. HERMANCE and T. F. EGAN

(Manuscript received December 12, 1957)

An amorphous brownish organic deposit has been found on non-arcing palladium contacts. It is shown that this material is produced when a number of metals, notably those of the palladium and platinum groups, are rubbed together in the presence of low concentrations of organic vapors. The deposit is given by a very wide variety of organic materials, and is thought first to involve chemisorption of the vapor on the metal, followed by frictional activation and polymerization of the chemisorbed layers to form the visible accumulation. The use of a gold alloy overlay on the palladium to minimize the troublesome effects of the polymer on contact performance is discussed.

I. HISTORICAL REVIEW

The reliability of telephone switching and the quality of speech transmission are vitally dependent on the effective performance of the hundreds of electrical relay contacts employed in the completion of a call. Despite all reasonable efforts to maintain clean operating conditions, the function of a contact may be impaired by a variety of contaminants, the sources of which are often far from obvious. The discovery and control of such contaminants has been a continuing task at the Laboratories for many years. In 1947, however, a new approach was made to the problem, with the employment of the plastic replica technique as a micro-analytical tool.[1] Briefly, the method consists of pressing the warmed contact into a clear thermoplastic sheet material such as Vinylite. When cooled, the resulting imprint reproduces faithfully all surface details of the contact. Even more important, foreign deposits are transferred in their original configuration to the transparent medium for microchemical study.

Through this medium of sampling and analysis, a systematic study was made of contact contamination as it occurs under the conditions
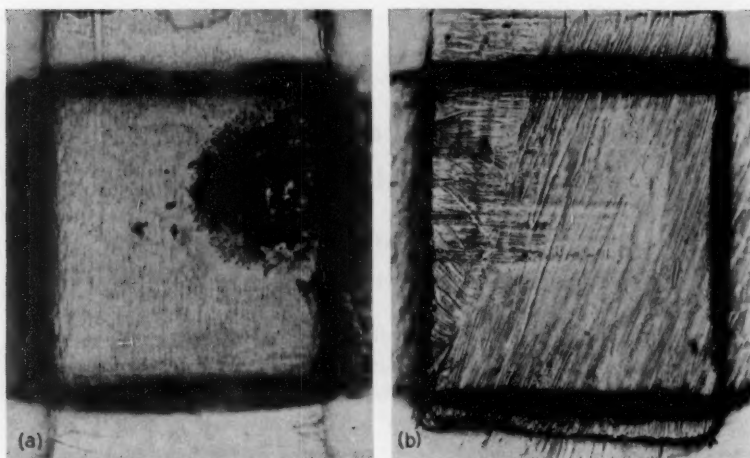
Fig. 1 — (a) Vinylite replica of a palladium relay contact, showing organic deposit formed from vapors of organic structural parts after 10⁶ closures; (b) replica of a silver relay contact after 10⁶ closures, showing complete absence of organic deposit. (Both magnified 75 times.)

of manufacture, installation and switchroom operation. This study led to the discovery of a wholly unexpected source of contamination, which appears to explain many of the mysterious transient "opens" frequently encountered in switching. It was found that a dark brownish, structureless deposit was invariably present on non-arcing palladium but never on silver contacts. Under the microscope, the deposit was always observed to accumulate around the area of contact impact and slide. Fig. 1(a) shows a Vinylite replica of a palladium relay contact contaminated with the brown deposit produced by about one million closures. Fig. 1(b) shows a replica of a silver contact operated under the same conditions. In a typical case, most of the deposit is found considerably outside the contact area as a dark powdery mass. Closer to the area of slide may be seen thin transluscent films which transmit a yellow to red light. These films are formed by impaction and smearing of the somewhat plastic deposit.

This contaminant could not be traced to any known atmospheric or shop source. Moreover, its quantity was undiminished on palladium contacts operating in the cleanest environments, virtually free from all air-borne contaminants. Significantly, too, this material was found on unwired contacts of operating relays but never on contacts which had not been subject to operation or vibration.

To summarize, this mysterious brown deposit appeared on palladium but not on silver contacts. Its occurrence was completely independent of such environmental factors as air-borne dusts and corrosive gases. Its formation required mechanical but not electrical operation of the contacts. It accumulated in the area immediately surrounding the region of contact slide. Arcing electrical discharges appeared to destroy such deposits, completely or in part.

These peculiar circumstances led to the view that the deposit must be a frictional product, involving either palladium and dust-free air or palladium and some vapor derived from the relay structure. Micro-chemical studies made on replicas taken in the field failed to disclose more than traces of palladium in the deposits. Thus, although not easily acceptable, the evidence pointed strongly to vapors emanating from the relay structure as the source of the frictional product.

To test this view, new relays with carefully cleaned palladium contacts were enclosed in sealed glass chambers, with the contacts unwired. The magnet coils were energized at one-second intervals. At the end of several hundred thousand operations, the chamber was opened and replicas were made of the contacts. A brown deposit, identical in all respects to that observed in the field, was found around the frictional areas of the contacts. Similar experiments with silver contacts produced no deposit.

A relay with palladium contacts was then constructed in which all organic components were replaced by glass in sheet or fibrous form. As an additional precaution, this relay was solvent-extracted and out-gassed by repeated baking and evacuation. When it was operated as before in the sealed container, no deposit appeared on the contacts (see Fig. 2). However, when pieces of phenol fiber such as are used in the normal relay were introduced into this container, a deposit again appeared when the relay was operated. Likewise, deposits were obtained when this inorganic relay was operated in the presence of pure vapors such as benzene or hexane (see Fig. 3). The conclusion thus was inescapable that the brown deposit is formed by traces of organic vapors emanating from the insulating materials of the relay structure.

The purely organic nature of the deposit was confirmed by micro-chemical tests. The product was found to be combustible at red heat, leaving no residue. On the replicas, the material quickly disappeared in a drop of concentrated sulphuric acid containing chromic acid, with the formation of bubbles, presumably CO_2.

The electrical behavior of the brown deposit on the contacts was studied by the use of microprobes. The nonconducting nature of the
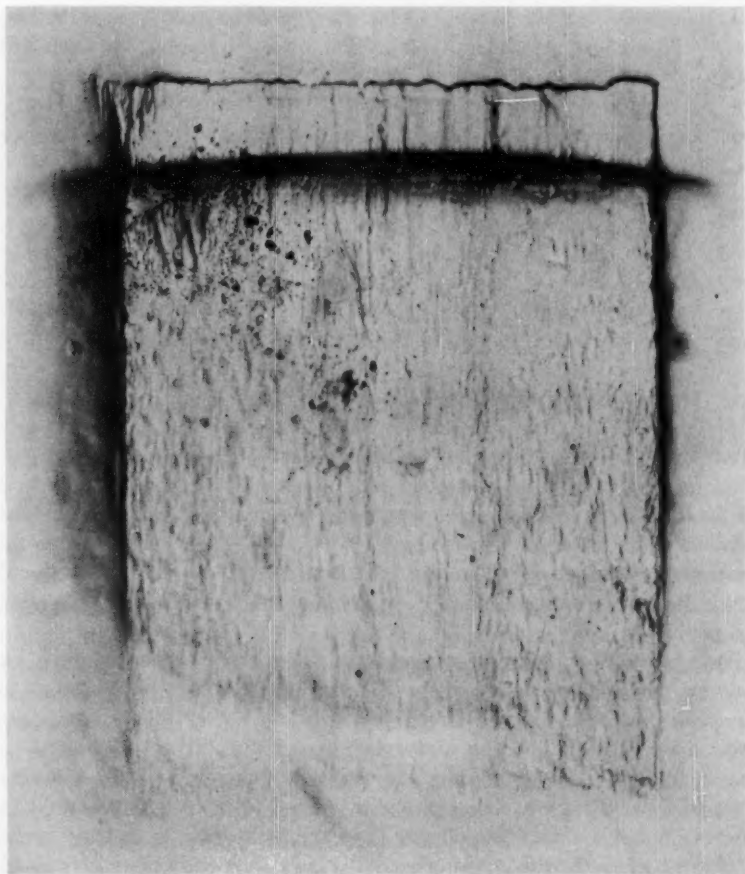
Fig. 2 — Replica of a palladium contact after 10^6 closures on a relay with all inorganic parts, showing only metallic wear particles. (Magnified 75 times.)

material was confirmed, but it was found that the deposits were quite fragile compared with fibers and other particles to which most persistent opens on operating equipment have been traced. Nevertheless, it was believed that many transient effects might be explained by this contaminant.

Confirmation of these suspected transient effects was obtained when an extensive laboratory study was made to compare relay performance with and without dust-excluding enclosures.[2] Test circuits were employed

in this study which were designed to detect all opens, even those of short duration. Despite the absence of dust, the relays in the tightest enclosures showed the most opens, and the only contaminant found on these contacts was the characteristic brown organic deposit. Enclosure of the relays evidently had trapped their organic vapors, accelerating the formation of the deposit.

II. PRODUCTION OF THE DEPOSIT IN LABORATORY DEVICES

Attention was next directed to reproducing the brown deposit in the laboratory by a simplified mechanism which would provide for
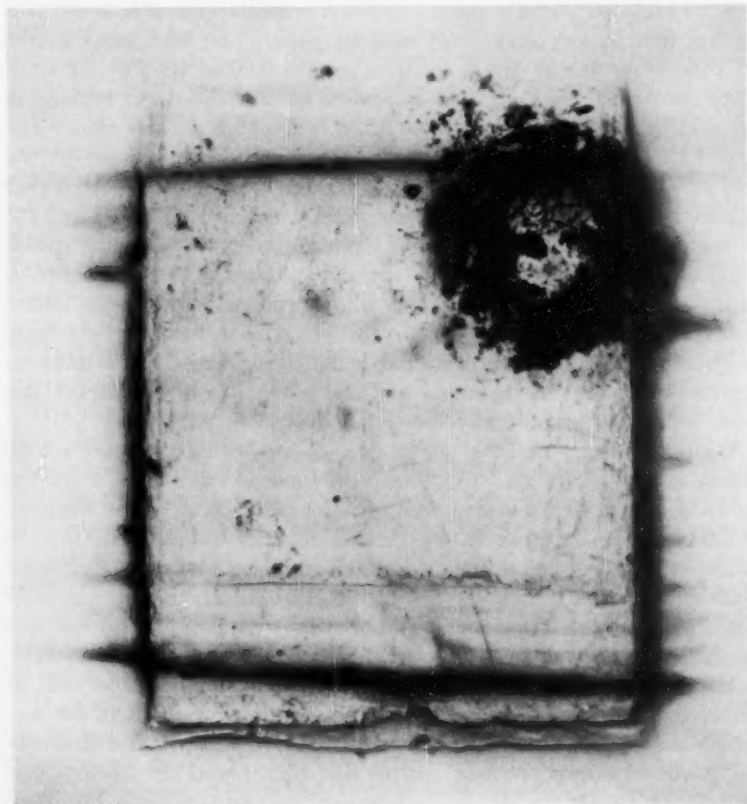
Fig. 3 — Replica of a palladium contact with all inorganic components, showing organic deposit from hexane vapors after 10^6 closures. (Magnified 75 times.)

better observation and more quantitative control. It was particularly desired to examine the behavior of metals other than palladium, and to produce a sufficient quantity of the organic material from known vapor sources to permit further study of its properties.

Closure of relay contacts involves both impact and slide. It was believed that impact played no part in producing the deposit, and this matter was settled by adjusting the springs of a relay so that the contacts slid over each other but never opened. The brown deposit was still produced in about the same quantity, indicating slide to be the essential motion.

This conclusion was verified by a simple device in which a palladium disc was rotated under a small ball fused onto the end of a 30-mil palladium wire in the presence of benzene vapor. The wire was soldered to a phosphor bronze spring and was bent to trail on the disc, at a low angle to reduce chatter. After several thousand rotations, a ridge of brown deposit was observable on either side of the circular track. The best results were obtained when the wiping speed did not exceed 10 cm/sec and the contact force was not greater than 35 gms. With higher speeds and contact load, wear was greatly accelerated and the yield of organic deposit fell off rapidly. Even at the most favorable speeds and loads, this device behaved erratically, at times producing the brown organic material copiously and at times producing mostly palladium wear products. It was believed that this behavior was caused by a high-frequency chatter inherent in the particular construction. It was interesting to note, however, that wear and the production of the organic deposit tended to be mutually exclusive effects.

On the rotating disc, the brown deposit was distributed over a long track, making it difficult to collect and to observe. Perfection of the rotating-disc device was therefore abandoned in favor of the development of one using reciprocating motion of small enough amplitude to confine the entire slide to the field of a microscope. The feature most attractive in such a device was the possibility that the rubbing elements could be so mounted that they could be readily detached for microscopic study, weighing and chemical examination of the friction product.

Exploratory devices of this type gave a fairly reproducible yield of brown deposit when the wipe velocity was held below about 10 cm/sec. Above this velocity, wear was again noticeable, with erratic and diminishing yields of brown product, just as had been found on the rotating-disc device.

Two machines using reciprocating motion were evolved. One is driven magnetically using 60-cycle alternating current which produces two

magnetic pulses per cycle and hence 120 oscillations per second of the slider (240 wipes per second). The other machine employs a motor-driven cam and lever arrangement for actuating the slider. Figs. 4, 5(a) and 5(b) show the general features of the magnetically driven device. The fixed friction member is a small plate, 0.5 cm × 1.0 cm, which may be recessed for retention of the friction products. This plate is held in a vise attached to the frame of the machine. The slider is a 30-mil wire, the end of which is bent sharply back on itself to provide a rounded sliding surface. The slider is clamped at the end of a phosphor bronze spring so that its tip is in line with the axis of the spring. The spring, which may be adjusted to the desired tension, is mounted on the end of the driving rod. No special finish was needed on either the slider or plate to obtain reproducible results, although polishing with levigated alumina facilitated microscopic study of the track by eliminating confusing scratches and other surface irregularities.

The driving rod is suspended in the frame by two beryllium-copper cantilever springs which provide the restoring force against the pull of the magnet. Early experiments showed that support of the driving rod in sleeve bearings was undesirable because of a tendency for it to wear and bind, even when lubricated with graphite. The driving rod terminates in a disc-shaped soft iron armature which is actuated by an E-shaped electro-magnet. The magnet assembly is attached to the frame in a sleeve clamp which permits adjustment of the armature gap. By this means and by regulation of the current, the stroke amplitude is controlled.

The frictional parts of the machine are enclosed in a tight-fitting cylindrical glass shell. A thin copper diaphragm is placed between the magnet and the armature, sealing off the rear of this chamber. Thus, the coil, the only part containing organic material, is on the outside. All parts enclosed by the chamber, except the actual friction members, are gold-plated to reduce reactivity to vapors of a chemically corrosive nature. Vapors to be tested may be passed through the chamber with any desired diluting gas, and inlet and outlet tubes for this purpose are provided. Alternately, liquids or solids to be vaporized may be placed inside the chamber in glass capsules.

The motion of the slider is simple harmonic, making 120 oscillations per second on the fixed plate, or 240 wipes. The average wipe velocity therefore will be $S \times N$, where S is the amplitude and N the number of wipes per second. The stroke amplitude is measured microscopically, with the travel of a fine scratch on the spring being observed. The actual velocity in simple harmonic motion will, of course, vary from zero at

Fig. 4 — Electromagnetically driven friction device.

GAS:
INLET OUTLET

SLIDER
CLAMP
SCREW

VISE

SLIDER SPRING

CANTILEVER
SUPPORTING
SPRINGS

TUBULAR
FRAME

ARMATURE

COPPER
SEALING
DIAPHRAGM

DRIVING MAGNET

GAP
ADJUSTING
SCREW

COIL

MAGNET
CLAMPING
RING

DRIVING ROD

GLASS ENVELOPE

SLIDER
PLATE

SLIDER
WIRE

0   1   2
INCHES
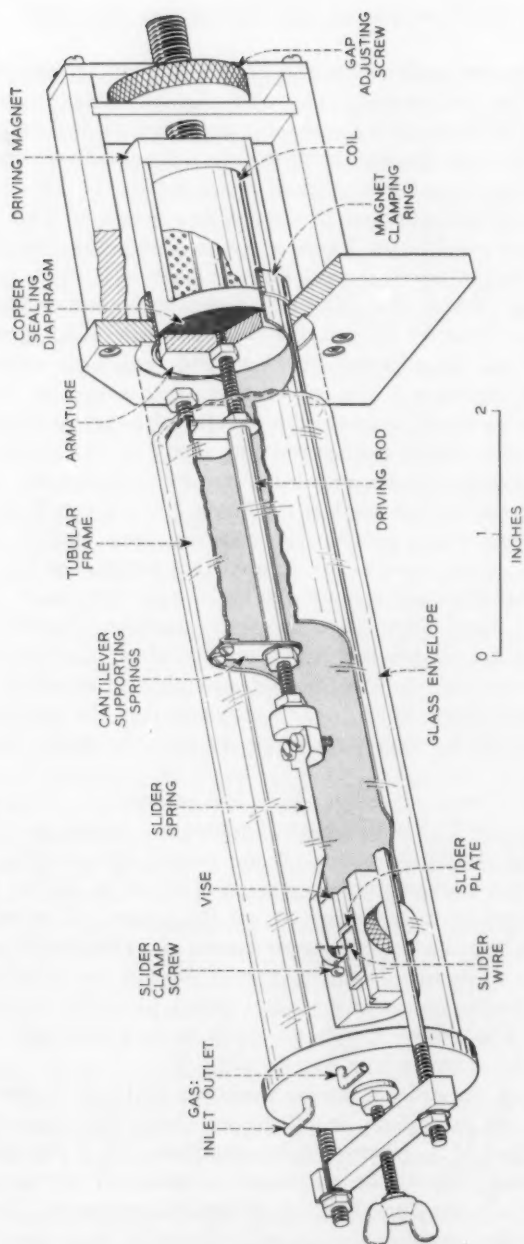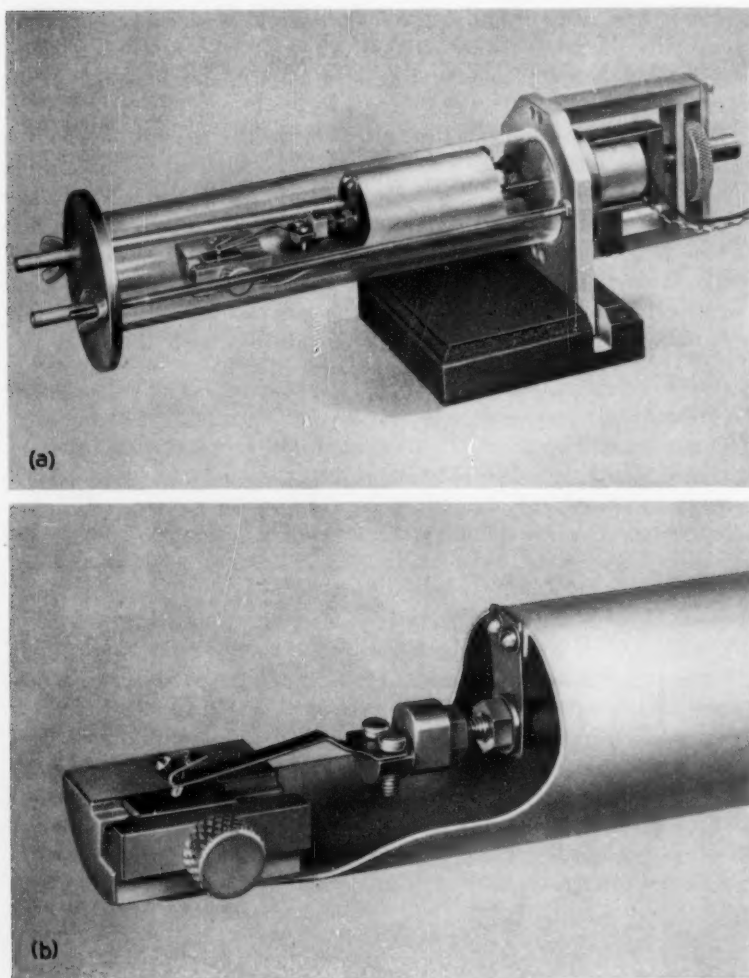
Fig. 5 — Close-up views of electromagnetically driven friction device.

the ends of the stroke to a maximum, given by $\pi/2\,(S \times N)$ or $1.57\,(S \times N)$, at the midpoint. Thus, with the amplitude of 170 microns commonly used, the maximum wipe velocity would be 6.41 cm/sec.

To measure yields of organic deposit, the plate and slider wire are weighed together on a microbalance before and after each run. The

gain in weight, after drying at 100°C, is taken to be the organic product yield. After the products are removed, the plate and slider are cleaned in redistilled methyl alcohol and again dried. The loss in weight is the wear.

Mechanically, this magnetically driven device is more susceptible to small variations in stroke amplitude than the more rugged and precise cam-driven machine. It was devised to obtain workable quantities of deposit quickly from various combinations of known vapors and metals. Molecules of the organic vapor reach the metal surface by diffusion through the diluting gas. This process is relatively slow when the vapor concentrations are low, as in the case of those emanating from relay components. Therefore it is possible, with the short rest
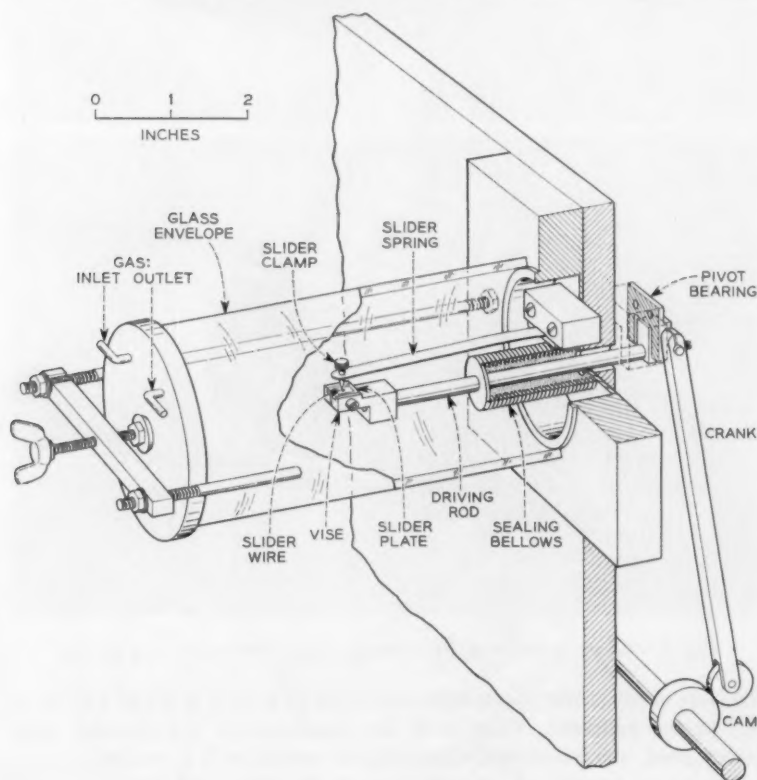
Fig. 6 — Cam-driven friction device.

period afforded by 120 oscillations per second and a low vapor concentration, that not enough adsorption will take place between each wipe to produce the needed film and little or no deposit may result. For this reason, the experiments with the magnetic device are made using high vapor concentrations. A sufficient supply of vapor molecules for rapid saturation even at high operation rates thus is assured.

The cam-driven machine, on the other hand, operates at frequencies up to 6 cycles per second and thus comes within the range of relay test cycling. It is intended for more precise studies, particularly those involving low vapor concentrations. It has been used, for example, to evaluate deposits formed from vapors emanating from materials used in relays and associated equipment and their proposed substitutes.

Fig. 6 shows the details of the cam machine. The friction elements are essentially the same as in the magnetic device, and are likewise enclosed in a glass envelope. In this case, however, the plate vise is mounted on a stiff rod which pivots at the back of the machine. The rod is moved by a lever arm and is spring-loaded against a motor-driven cam. The opening in the back plate through which the rod passes is sealed off by a metal bellows which surrounds the rod. In operation, the rod oscillates through a small vertical arc determined by the cam eccentricity and the length of the lever arm.

The slider, again a 30-mil looped wire, is held in a small chuck at the end of a nickel-silver cantilever spring mounted on the back plate and above the driving rod. As the rod moves up and down against this spring, slip occurs between the plate and slider in much the same manner as between the contacts of a relay. The motion is still essentially simple harmonic, but there will be a difference in contact force between the top and bottom positions of the rod. On the device in use, the force at the bottom of the stroke is about 70 per cent of that at the top. The stroke is set for 170 microns and the slide velocity will therefore depend on the frequency of operation. At 6 cycles per second the average velocity is 0.2 cm per second.

III. INITIAL EXPERIMENTS WITH BENZENE VAPOR

Benzene was used as the vapor source for the first series of observations for several reasons: (1) the deposit formed from it resembled closely the one found on contacts of relays in the field; (2) the benzene ring is a characteristic part of the structure of most of the vapors given off by relay components; (3) pure benzene is easily obtained; (4) its high vapor pressure permits a wide range of vapor concentrations by dilution of benzene saturated air.
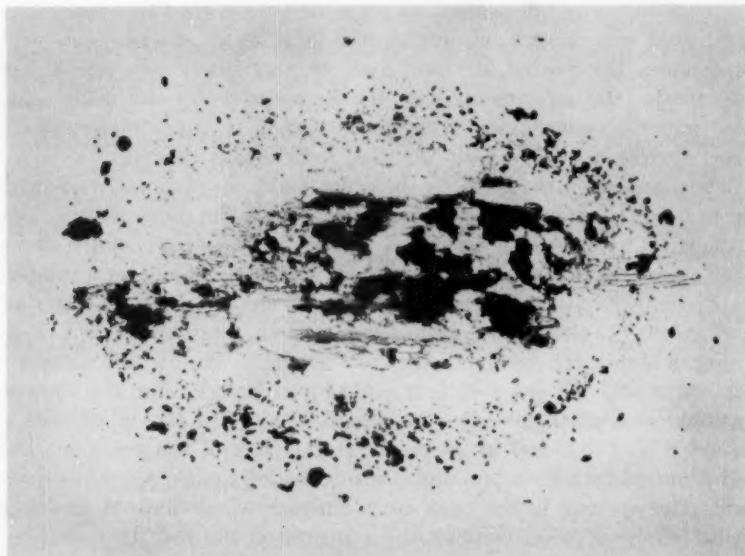
Fig. 7 — Replica of palladium plate, showing organic deposit formed after one minute in benzene vapor with 120-cycle machine. (Magnified 400 times.)

Air was first passed through a column of benzene by means of a fritted glass bubbler, then through the 120-cycle magnetic device with the stroke adjusted to 170 microns. The sliding elements were palladium and the spring force was 30 grams. At the end of ten seconds (2400 wipes) a replica of the plate was made and examined microscopically. A shallow wear area was found to contain numerous furrows and scratches, obviously the composite pattern formed by lateral variations in the slider position. Palladium wear fragments were scattered in and around the wear area, together with amber to dark brown transluscent flakes of organic materials. At the bottom of many of the furrows there was a thin, continuous film of organic deposit. After one minute (Fig. 7), the wear had not increased appreciably, but the low areas were now almost filled with the organic material. From this point on, successive observations disclosed no further wear, the organic friction product apparently providing effective lubrication. At the end of five hours (4,320,000 wipes) the product had accumulated around the slide area as a brown powdery pile. The yield amounted to about 35 micrograms.

Another run was started, but after five minutes the machine was

stopped and the wear products were removed without disturbing the slider. It was started again and operated for five hours, after which all of the friction product was removed and ashed. Only a trace (less than 0.1 microgram) of palladium was found, confirming the belief that wear is negligible once an appreciable organic film is formed.

In view of the marked suppression of wear attending generation of the organic product, a comparison was made of the amount of wear produced in the absence of organic vapors under the same conditions. Figs. 8(a) and 8(b) show the result after 4,320,000 wipes. A deep pit has worn in the plate, with a voluminous pile of finely divided palladium surrounding it. Figs. 9(a) and 9(b) show the result of the benzene vapor run with the organic deposit surrounding the track, and with this deposit removed to show the small wear area.

Table I compares quantitatively the wear obtained on three benzene
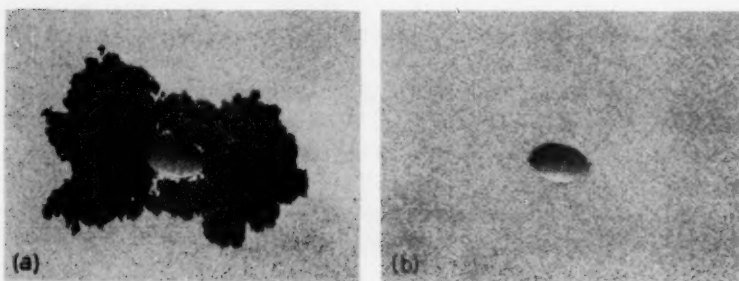
Fig. 8 — (a) Palladium wear products generated in the absence of organic vapors after about 4×10^6 wipes; (b) deposit removed, showing deep pit in absence of lubrication by the organic deposit. (Magnified 25 times.)
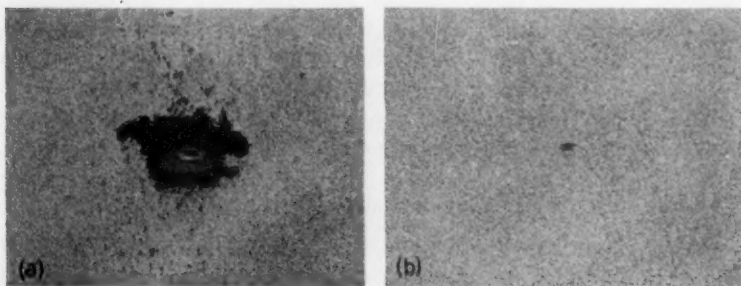
Fig. 9 — (a) Organic deposit formed from benzene vapors after 4×10^6 wipes; (b) deposit removed, showing greatly reduced wear due to lubrication by the vapors and deposit. (Magnified 25 times.)

TABLE I — COMPARISON OF WEAR — WITH AND WITHOUT BENZENE VAPOR

$(4 \times 10^6$ Wipes, 170-Micron Stroke, 30-gm Force, 120 Cycles/Sec)

Experiment Number	Gas Phase	Yield of Organic Friction Products	Average	Yield of Palladium Wear Product	Average
		Micrograms		*Micrograms*	
1	air	—		103	
2	air	—		81	100
3	air	—		117	
4	benzene sat. air	46		4	
5	benzene sat. air	32	44	4	3.5
6	benzene sat. air	35		2.5	

runs with the wear on three "dry" runs. In the benzene runs, no attempt was made to remove the initial wear products and the total wear was obtained by transferring the friction product to a weighed platinum capsule, igniting the organic material and re-weighing to obtain the palladium residue.

A series of runs was made in benzene saturated air on both the magnetic and the cam-driven machines. The results are presented in Table II. The greater reliability of the cam machine is evident. It is interesting, however, that the average yield with these two devices is substantially the same in high benzene concentration, even though one has a slide velocity and frequency twenty times that of the other.

IV. YIELD OF DEPOSIT AS A FUNCTION OF SLIDER LOAD

A series of runs was made at several spring loads on the 120-cycle machine with other factors held constant. Fig. 10 shows the results, with yield of deposit plotted against the square root of the slider load.

The yield appears to be proportional to the square root of the load, and the increase in track width due to plastic deformation between the curved slider and the plate would also be proportional to the square root of the load. Hence, it would appear that the load regulates the yield through regulation of the track area.

V. MONOLAYERS OF BENZENE TRANSFORMED PER WIPE

If we take 35 micrograms (from Table II) as the average yield of friction product for 4,320,000 wipes, with about 4 micrograms consisting of palladium wear particles, the organic deposit formed per wipe will be about $(35 - 4)/4 \times 10^6$, or 8×10^{-6} micrograms.

TABLE II — YIELD REPRODUCIBILITY — MAGNETIC AND CAM MACHINES

(4 × 10⁶ Slides, Benzene Saturated Air, 170-Micron Stroke)

Yield in Micrograms

Magnetic — 120 Cycles/Sec 30-gm Spring Force		Cam — 6 Cycles/Sec 20–30-gm Spring Force
44	40	39
46	50	36
30	36	29
38	36	36
30	41	38
32	38	30
Average = 38.4		Average = 34.7
Average deviation = 4.8		Average deviation = 3.3

R. W. Wilson[3] has determined the track width for a gold wire slider on gold. Wilson's data include results of direct microscopic measurements, calculations from contact resistance and values calculated from Hertz and Meyer's equations at 30 grams force. All of these methods agree rather well, giving a track width for annealed gold of 5×10^{-2} mm. From this value, R. L. Barns[4] calculates the track width for work-
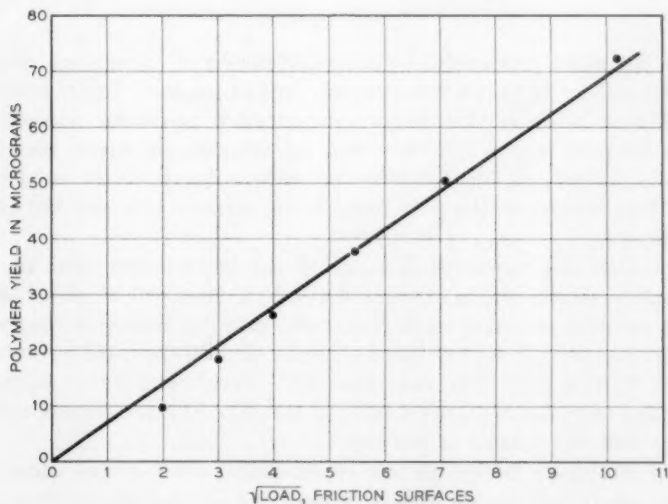
Fig. 10 — Polymer yield versus load. (Palladium on palladium; 4 × 10⁶ wipes in benzene-saturated air; 170-micron stroke.)

hardened palladium, using the relation:

$$\frac{d_{Au}}{d_{Pd}} = \sqrt{\frac{H_{Pd}}{H_{Au}}}$$

where d_{Au} and d_{Pd} are the track widths, respectively, for gold and palladium and H_{Au} and H_{Pd} are the hardness values in kg/mm² for gold and palladium, respectively. The track width thus obtained for palladium is 2.4×10^{-3} cm. With a 170-micron stroke, the track area will be 4×10^{-5} cm².

The 8×10^{-6} micrograms of deposit formed per wipe corresponds to 1×10^{-13} moles of benzene, which in turn is equivalent to 6×10^{10} molecules of benzene. Dividing this value by twice the wiped area gives about 7.5×10^{14} molecules/cm² formed per wipe.

Twice the wiped area is taken on the assumption that the surfaces of both the slider and the plate generate equal quantities of polymer. This would be the case if the relatively small active area of the slider were instantaneously regenerated as it traverses the plate.

A monolayer of benzene on palladium contains about 7×10^{14} molecules/cm². Hence, there has been converted one monolayer per wipe on each surface.

VI. VAPOR CONCENTRATION VERSUS YIELD OF PRODUCT

As discussed previously, high concentrations of vapors are used in the laboratory to insure a maximum yield of product. Under practical conditions, however, the vapor concentration producing deposits on relay contacts is probably very low. An attempt was made, therefore, to find the critical concentration of benzene below which no deposit would be formed on the palladium sliding surfaces at 6 and 120 cycles per second.

For obtaining measured dilutions of the benzene-saturated air, the technique shown in Fig. 11 was devised. A glass cell of about 50 cc capacity with platinum electrodes sealed into the bottom is connected to the metered air supply line by means of a 10-cm capillary tube of about 0.2-mm bore. The electrodes are covered with dilute sulphuric acid and over this is placed a layer of benzene. The two liquids occupy about half the volume of the cell.

The free space in the cell will be saturated with benzene vapor and its transport into the air supply through the fine capillary will be negligible until gas pressure is generated in the cell by electrolysis of the dilute sulphuric acid. By this means, the small required amount of

benzene vapor is forced continuously through the capillary into the main air supply stream. The volume of vapor is thus metered by the current supplied, in accordance with Faraday's Law. The minute amount of hydrogen introduced into the air stream has no effect on the production of the organic deposit, and the rate of removal of the vapor from the cell is so slow that benzene saturation in the cell is essentially maintained.

Runs were made at 6 and 120 cycles per second using 30 grams spring force and 4×10^6 slides at several concentrations. The yields of deposit were plotted against the vapor concentration at the two slide frequencies and are shown in Fig. 12. The values near the critical concentrations are approximations, since estimation of these low yields is subject to considerable error and the cut-off point must be judged from microscopic evidence alone. The latter observation is aided, however, by the abrupt increase in wear and change in the track texture as the organic product disappears. Even with these limitations, it seems clear that with the lower wipe frequency the critical concentration is lowered. At 240 wipes per second, a concentration of about 35 parts per million is re-
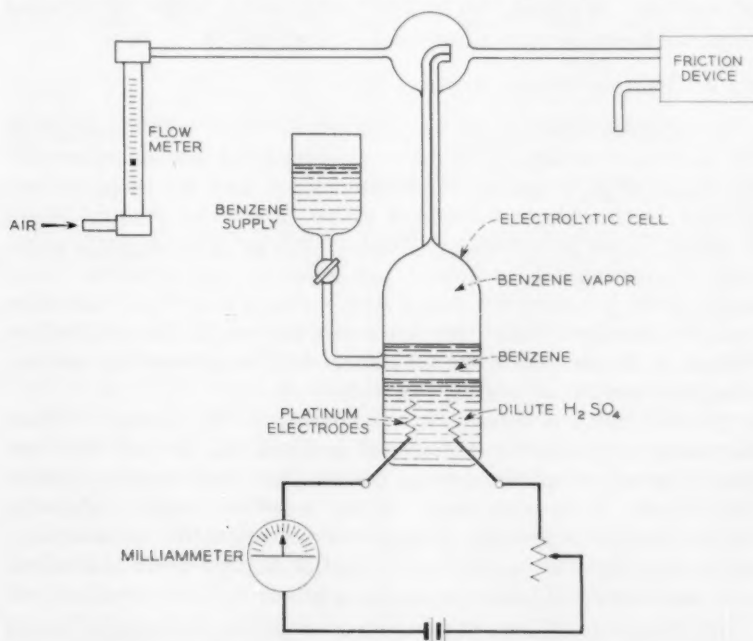
Fig. 11 — Device for obtaining metered dilutions of benzene vapor.
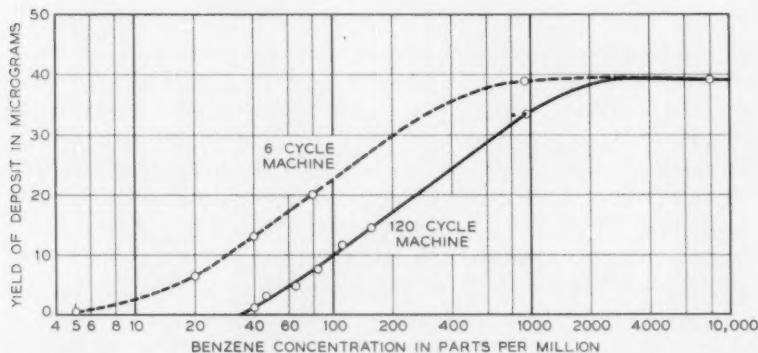
Fig. 12 — Yield of benzene deposit as a function of vapor concentration and wipe frequency.

quired for the appearance of the deposit. At 12 wipes per second, about 5 ppm are needed.* At the relatively infrequent rate of relay operation in telephone switching, the critical concentration might be expected to be still lower, perhaps a fractional part per million.

VII. PROPERTIES OF THE BENZENE PRODUCT

By extended operation of the 120-cycle device in benzene-saturated air, enough of the brown product was obtained to permit a closer examination of its properties. The material is a dark brown amorphous powder which transmits a reddish amber light under the microscope. It readily takes on a strong electrostatic charge. The powder is somewhat plastic and can be formed into pellets at high pressures. When such a pellet is heated in a closed tube, it does not melt but carbonizes with little change of form to produce a coke-like residue. An oily distillate collects at the cool end of the tube, suggesting the presence of considerable hydrogen in the original material.

A wide variety of solvents has been tried on the benzene product, the results of which are summarized in Table III. No solubility has been observed except with strong organic bases such as ethanolamine and pyridine. Even with these, solution is never complete. When the soluble portion in pyridine or morpholine is run into a silica-impregnated paper from a capillary tube a radial chromatogram is obtained with several colored bands, indicating a mixture of several substances.

---

* P. Kisliuk,[5] using gas (hydrogen) pressure increase as an indicator of polymer production in an evacuated system, obtains polymer at a benzene concentration of about 4 ppm at 2.6 cycles per second, which is in general agreement with these conclusions.

TABLE III — EFFECT OF VARIOUS SOLVENTS ON THE
BENZENE PRODUCT

Solvent	Action
Aliphatic hydrocarbons (petroleum fractions) Cyclic hydrocarbons (toluene, xylene, cymene, tetralin, etc.) Chlorinated hydrocarbons — aliphatic and aromatic Alcohols — methyl to amyl Ethers — dimethyl, diethyl, dipropyl Ketones — dimethyl, diethyl, methyl butyl Esters — ethyl acetate, methyl salicylate, amyl acetate, dibutylphthalate Acids — glacial acetic Aniline, dimethyl aniline, dodecylamine, alpha-methylbenzylamine Acetonitrile, dioxane, nitropropane, dihydropyran Furfuryl alcohol, dimethylformamide, dichloroisopropyl ether	No Action
Sulphuric acid 75%, zinc chloride (conc. aqueous) Sodium thiocyanate (conc. aqueous), sodium hydroxide, 20%	No Action
Nitrobenzene, diethylene glycol monoethyl ether, ortho-cresol	Slight swelling of particles under microscope
Ethanolamine, isopropanolamine	Largely soluble
Pyridine, quinoline, butylamine	Largely soluble
Ethylene diamine, propylene diamine	Partly soluble
Morpholine	Partly soluble
Conc. sulphuric acid + chromic acid	Slowly disperses and disappears with formation of CO_2 bubbles
50% aqueous potassium hydroxide, triglycol, octoic acid, cresols; all in sealed tubes, heated to 200°C.	Product swells slightly but no evidence of solution

The carbon-hydrogen atomic ratio was determined on two 500-microgram samples and found to be 1.36 and 1.18, giving an average of 1.27. The spread between these results probably reflects some non-uniformity in the sample material as well as the general difficulty of working with such small quantities. In any event, it is evident that, in the conversion from the benzene vapor to the brown deposit, much of the hydrogen is retained.

P. Kisliuk,[5] by measuring the pressure increase in a closed system containing benzene vapor and sliding palladium wires, has determined the hydrogen released during polymer formation. He concludes that the polymer retains approximately 65 per cent of the benzene hydrogen.

An attempt was made to obtain an infrared spectrum of the benzene product. The small quantity available together with its insolubility

created difficulties, but there was evidence of aromatic rings, alkene and carbonyl groups linked together to form long chain molecules. Electron diffraction studies also indicate an amorphous material of high molecular weight.

In reviewing these properties, one is inclined to regard the organic deposit as a complex mixture of high molecular weight, cross-linked materials. The infrared indication of carbonyl groups, the strong color of the deposit, and the fact that only basic solvents have any action on it, strongly suggest that, in the conversion from benzene vapor, oxygen has entered into combination. Perhaps cross-linkage is through oxygen-containing groups such as aldehyde or carboxyl, and, by attacking these, the basic solvents achieve splitting and solution. Since the deposits resemble mixed polymers in their properties, it is proposed to term them "frictional polymers".

VIII. BENZENE FRICTIONAL POLYMER FORMED WTIH SLIDING SURFACES OTHER THAN PALLADIUM

Various sliding surfaces in benzene vapor were substituted for palladium in the 120-cycle device. Table IV presents the results. It is apparent that generation of the organic deposit is a phenomenon in which the platinum and palladium metals are especially active, although it is not exclusively confined to these. Molybdenum, tantalum and chromium, for example, give weighable quantities of deposit. On the other hand, alloys with only small percentages of platinum or palladium still produce the deposit, in reduced quantities. Silver on silver produces no detectable organic film, nor, curiously, is wear observed. Even when it is used against palladium the deposit is negligible, probably because of frictional transfer of silver to the palladium surface at the outset of the experiment. Gold produces a very small quantity of product, the nature of which appears to be identical with the deposit formed with palladium.

If the catalytic properties of the platinum and palladium metals are important in production of the polymer, one might also expect nickel to be active. In air, the only frictional product is nickel oxide. To prevent oxidation, benzene-saturated hydrogen (which gives polymer with palladium) was used, but only wear was observed.

When the slide is between quartz or glass and palladium, polymer is still produced (Fig. 13). Apparently metal-on-metal slide is not necessary. Although it may be argued that traces of palladium transfer to the quartz, no indication of this was obtained by microscopic or microchemical means. Attempts were made to use other non-metallic sliders

Table IV — Behavior of Various Sliding Surfaces in Benzene Vapor

(4 × 10⁶ Wipes, 170-Micron Stroke, 120 Cycles/Sec, 30-gm Force)
(Except Where Otherwise Indicated, in Benzene-Saturated Air)

Sliding System*	Weight of Frictional Product (Micrograms)	Description of Frictional Product
Pt/Pt	40	Dark brown polymer, powder
Pd/Pd	38	Dark brown polymer, powder
Pd/Pd (Benzene-sat. H₂)	35	Dark brown polymer, powder
Ru/Ru	22	Dark brown polymer, powder
Pt 6%, Au 69%, Ag 25%/Pd	20	Brown polymer, powder
Pd 25%, Au 75%/same	15	Dark brown polymer, powder
Ta/Ta	14	Gray polymer, powder
Pd 60%, Cu 40% ("unordered")/ same	13	Dark brown polymer, powder
Quartz/Pd	12†	Brown polymer, powder
Au 91.7%, Ag 8.3%/Pd	12	Brown polymer, powder
Rh/Rh	10	Light brown polymer, powder
Pd 60%, Cu 40% ("ordered")/same	9	Dark brown polymer, powder
Au 70%, Ag 30%/Pd	8	Dark brown polymer, smeary powder
Mo/Mo	8	Light brown polymer, powder
Cr/Cr	8	Mixed tan and light brown polymer
Au/Pd	5–8	Dark brown polymer, smear
Pd 5%, Au 95%/same	5†	Brown polymer, powder
Pt 6%, Au 69%, Ag 25%/same	5†	Brown polymer, powder
Pd/Quartz	3–5†	Brown polymer, powder
Au/Au	1–3†	Dark brown, smeary polymer
Ag/Pd	1–3†	Dark brown, smeary polymer
Ni/Ni	None	Yellow brown oxide
Ni/Ni (Benzene-sat. H₂)	None	Metallic wear products
Ag/Ag	None	Little wear
Cu/Cu (Benzene-sat. H₂)	None	Some metallic wear
Quartz/Quartz	None	Fine quartz wear powder
Fe/Fe (Benzene-sat. H₂)	None	Wear products
V/V	None	Brown oxide
W/W	None	Oxide formed
Zn/Zn (Benzene-sat. H₂)	None	Wear products

* The first member of the pair is the slider.
† Estimated microscopically.

such as nylon, Vinylite, and cellulose on palladium, but no polymer was obtained.

IX. ATTEMPTS TO INHIBIT FRICTIONAL POLYMER FORMATION

Again assuming catalysis is involved, it would be reasonable to expect that the alloying of small amounts of certain metals with the palladium might produce a poisoning effect and inhibit the generation of the poly--

mer. Alloys containing arsenic, lead, antimony, phosphorus, bismuth and mercury were tested, but none showed any useful reduction in the quantity of polymer produced (Table V). These tests were carried out in the 120-cycle machine, both in benzene-saturated air and in air saturated with ethylene glycol dibutyl ether vapor. The latter was used on the assumption that agents which fail to inhibit in benzene-saturated air might work at low vapor concentrations. Ethylene glycol dibutyl ether gives a vapor concentration of about 120 ppm at room temperature and yields about 14 micrograms of a dark brown powdery polymer with 4×10^6 wipes (palladium on palladium).

Polymer formation was completely prevented when an atmosphere of 50 ppm of mercury was maintained in the chamber. However, under these conditions, the sliding surfaces contained a high concentration of mercury, which practically eliminated palladium in the frictional process.

Pre-treatment of the palladium surfaces with various films was unsuccessful (Table V). This in not surprising, since most surface coatings such as sulphide or oxide would be worn through quickly by the sliding action. Benzoquinone and sulphur coatings were applied with the idea of saturating the air immediately around the wiped area with their vapors. Unsuccessful attempts were also made to prevent polymer formation by contaminating the palladium surfaces with lubricants such as silicone greases, petrolatum and graphite.

A number of gases and vapors were also examined for possible inhibiting effects. Such agents were chosen either for possible poisoning action on the palladium surfaces or blocking effect on the growth of the organic molecules to the polymer stage.

Fig. 13 — Polymer formed on 120-cycle machine by palladium slider on fused quartz, with reflected light. (Magnified 50 times.)

TABLE V – ATTEMPTS TO INHIBIT FRICTIONAL POLYMER PRODUCTION

$(4 \times 10^6$ Wipes, 30-gm Force, 120 Cycles/Sec)

A. By Alloying of Certain Metals With Palladium

Palladium Alloyed with: (Both Sliding Members)	Polymer Yield, Air Saturated with:	
	Ethylene glycol dibutyl ether	Benzene
Unalloyed Palladium	14 micrograms	38 micrograms
Lead 1%		
Lead 15%		
Antimony 1%		
Antimony 15%	All yields were above	All yields were above
Arsenic 1%	10 micrograms*	20 micrograms
Phosphorus 0.8%		
Bismuth 4%		
Mercury 0.1%		

B. By Pre-treatment of Sliding Palladium Surfaces

Pre-treatment	Polymer Yield in Air Saturated with:	
	Ethylene glycol dibutyl ether	Benzene
Untreated Palladium	14 micrograms	38 micrograms
Electrolytic sulphiding + sulphur coating		
Benzoquinone deposited from 5% acetone solution	All yields were above	All yields were above
Oxidation of surfaces	10 micrograms*	20 micrograms
Lubricant coating — graphite, silicone grease, petrolatum		

* If the yield was substantially that of the pure palladium run, judged visually, it was not weighed, since then there was no significant inhibition.

Table VI summarizes the results. The tests were run with four polymer forming vapors — benzene, tetrahydronaphthalene, triethylbenzene and ethylene glycol dibutyl ether — to provide a range of vapor concentrations. Aqueous ammonia, methyl alcohol and carbon tetrachloride, all in high concentrations, displayed an inhibiting action when combined with the lower concentrations of polymer producing vapors; they also reduced the yield with benzene saturated air. These materials, which form no polymer by themselves, probably occupy portions of the palladium surface by virture of their high concentration and therefore cannot be regarded as true catalyst poisons.

Hydrogen, used as the carrier gas, showed some inhibiting action, but only on the low concentrations of polymer formers. This was somewhat surprising in view of the known affinity of hydrogen for palladium.

TABLE VI — INHIBITION BY INTRODUCTION OF CERTAIN VAPORS

Inhibiting Agent	Polymer Yield (Micrograms) from Air Saturated with:			
	Ethylene glycol dibutyl ether 120 ppm	Triethylbenzene 260 ppm	Tetrahydronaphthalene 450 ppm	Benzene 1.2 × 10⁴ ppm
None	14	31	32	38
Water	>10†			>20
1% Carbon monoxide	>10			>20
1% Nitric oxide	>10	>20	>20	>20
100 ppm Hydrogen sulphide	>10			>20
* Carbon disulphide	>10			>20
* Thiophene	>10			>20
* Triethylamine	>10			>20
* Carbon tetrachloride	Trace	Trace	Trace	11, 15
* Methyl alcohol	None		Trace	11
* Aqueous ammonia (conc.)	None	None	Trace	7, 5
Hydrogen, replacing air	7, 7	10	>20	>20

* Capsules of inhibitor and polymer producing agent were exposed together in the closed chamber of the machine.

† If the yield was substantially that of the run without inhibiting agents, judged visually, it was not weighed, since then there was no significant inhibition. All of the values were judged visually.

X. FRICTIONAL POLYMER FROM VARIOUS ORGANIC MATERIALS

Thus far, benzene has been used almost exclusively as the vapor source in the laboratory while the various physical factors were being studied. The investigation would be grossly incomplete, however, without attempts to establish the range of organic materials which produce the organic deposit and to observe how vapor composition affects its yield and properties. To this end, a variety of vapors were tested in the 120-cycle machine. In most cases, air was saturated with the vapor by the passing of fine bubbles through the liquid. Fixed gases were metered to give a concentration in air of several hundred parts per million. Occasionally it was more convenient, particularly when the material was a solid, to place it in a capsule inside the glass envelope. At the end of each run the friction product was weighed and examined chemically and microscopically.

The wide range of organic compounds, many highly stable chemically, which produce a deposit is surprising. Except for a few simple materials such as methane and ethane and some of their derivatives, the only requirement seems to be that the substance have high enough vapor pressure to produce a concentration of 50 ppm or more. All of the products are colored, ranging from light brown to almost black. They are usually powdery, amorphous materials, although exceptions do occur

TABLE VII — FRICTIONAL POLYMER FROM VARIOUS ORGANIC COMPOUNDS*

(Pd/Pd, 120 Cycles, 30-gm Load, 170-Micron Stroke, 4×10^6 Wipes)

Group A — No Polymer Formed	
Methane ⎫ down to −60°C Ethane ⎬ Propane ⎭ Methyl alcohol	Formaldehyde Carbon tetrachloride Methylamine Ethylene diamine

Group B — Trace to 5 Micrograms	
Propane (−60°C) Butane (down to −60°C) Ethylene	Ethyl alcohol Ethyl chloride Ethyl ether

Group C — 5 to 19 Micrograms	
Pentane Hexane Propene Butene Butadiene Propanol Butanol Pentanol n-Hexanol 2-Hexanol Acetaldehyde Propionaldehyde Acetone Methylethylketone 2-Butanone 2,5-Hexanedione Acetonitrile Butyronitrile Hexanenitrile Hexachlorbenzene	Propylamine Butylamine Amylamine Hexylamine Triethylamine Pentadecafluorotriethylamine Butylchloride Hexachlorbutadiene n-Nitrobutane Cyclohexanol Benzyl alcohol Aniline Hydroxybenzyl alcohol Monochlorbenzene Thiophene Pyridine Dioxane Phenol Morpholine o-Cresol

Group E — 20 to 40 Micrograms	
Acetylene Benzene Toluene Xylene Naphthalene Cyclohexene	Cyclohexane Crotonaldehyde Butyraldehyde Cyclohexanone 5-Hexen-2-one Limonene

Group F — Over 40 Micrograms	
Acrolein Acrylonitrile	Styrene Benzaldehyde

* The highest grade chemicals commercially available were used without further purification.

TABLE VIII — FRICTIONAL POLYMER FROM STRAIGHT CHAIN COMPOUNDS

(4 × 10⁶ Slides, Pd/Pd, in Air Containing High Concentrations of Vapor; Yield in Moles ×10⁻⁵ in Lower Right-hand Corner)

(Yields for Propane and Butane Given at −60°C and +20°C)

CLASS	C_1	C_2	C_3	C_4	C_5	C_6
SATURATED HYDRO-CARBONS	METHANE 0	ETHANE 0	(+20)0 PROPANE (−60)tr	(+20)tr BUTANE (−60)0.7	PENTANE 1.4	HEXANE 1.9
SATURATED ALCOHOLS	METHANOL 0	ETHANOL 0.6	PROPANOL 1.3	BUTANOL 1.5	PENTANOL 1.9	HEXANOL 2.1
SATURATED AMINES	METHYL-AMINE 0	ETHYL-AMINE tr	PROPYL-AMINE 1.1	BUTYL-AMINE 1.0	PENTYL-AMINE 1.1	HEXYLAMINE 0.7
SATURATED ALDEHYDES	METHANAL 0	ETHANAL 1.8	PROPANAL 2.0	BUTANAL 2.7		
OLEFINS		ETHYLENE 0.2	PROPYLENE 1.4	BUTYLENE 0.9		HEXENE 2.1
OLEFIN ALDEHYDES			ACROLEIN 10.0	CROTONALDE-HYDE 3.8		
ACETYLENE		ACETYLENE 13.8				

in which smeary products or peculiar fibrous structures are produced. In the latter case, needle-like outgrowths are sometimes encountered which suggest the existence of some sort of molecular ordering. Examination in polarized light, however, has never revealed any optical activity in these structures. Table VII lists all of the compounds studied, grouped on the basis of polymer yield in micrograms.

Table VIII compares the polymer yield for a number of straight chain compounds ranging from C_1 to C_6. The yield is reported in moles × 10⁻⁵ for better comparison of molecules transformed. The saturated hydrocarbons form an interesting series. Methane and ethane, boiling at −161° and −88°C, respectively, produce no deposit at temperatures

as low as $-60°C$. Propane (boiling point $-42°C$) gives a trace of deposit at $-60°$ but none at room temperature, while butane (b.p. $+1°C$) yields a weighable deposit at $-60°$ and a trace at room temperature. Pentane and hexane, both boiling above room temperature, give, respectively, 1.4 and 1.9×10^{-5} moles at room temperature. From this series it would appear that the deposit yield is influenced both by the length of the hydrocarbon chain and the difference between its boiling point and the machine's operating temperature. The importance of this temperature differential would be understandable if it is assumed that the saturated hydrocarbons are held to the palladium only by relatively weak adsorptive forces.

The dependence of the deposit on chain length is better demonstrated by the alcohol series, all of which boil above room temperature. There is a regularly increasing yield from ethanol to hexanol. Methanol, the single carbon member, gives no deposit.

The introduction of carbonyl oxygen, especially in a terminal position on the molecule (aldehyde), gives a darker product and markedly increases the yield as well as its solubility. Ethanal, for example, gives three times as much product as ethanol, and the product is dispersible in aqueous ammonia, methanol, dimethylformamide and other solvents which have little effect on the deposits yielded by hydrocarbons or alcohols. The butanal product is fairly typical of the aldehydes. Under the microscope it disperses readily in a drop of 10 per cent aqueous ammonia, yielding a dark amber suspension which appears clear in direct transmitted light but scatters light in a Tyndall beam. When introduced into filter paper from a capillary, the suspensoid does not travel with the liquid front but separates as a ring between the point of entry of the solution and the wet boundary. The suspensoid may be precipitated from the ammonia by adding a crystal of sodium chloride or other salt. Thus, the butanal polymer suspension in ammonia behaves much as a typical colloid. This increased dispersability of the aldehyde and ketone products, particularly in aqueous ammonia, suggests that acidic oxygen groups may account for a considerable portion of the linkages. By attacking these polar links, basic solvents bring the material into colloidal dispersion. An infrared examination of the butanal product gave indication of both carbonyl and carboxyl groups.

By way of contrast, deposits formed from unsaturated hydrocarbons are usually light colored and highly insoluble, even in powerful basic solvents such as ethanolamine. Here the cross-linkage is probably mostly through the double bond, with relatively little oxygen in the molecule.

The olefin hydrocarbons (containing the double carbon-carbon bond) show an increase in yield over the corresponding saturated members. With ethene, this increase is not spectacular, a barely weighable amount being obtained at $-60°C$ and only a trace at room temperature. Although this gas is known to chemisorb on palladium,[5] its low boiling point ($-104°C$) may operate against this sorption to reduce the yield. The increase is most marked for the C_3 member, propene. Propane (b.p. $-42°C$) and propene (b.p. $-47°C$) have almost the same boiling points, so that this factor practically cancels out. However, propene yields 1.4×10^{-5} moles at room temperature while propane gives only a trace of polymer even at $-60°C$, clearly indicating the influence of the double bond. The difference becomes less pronounced when butane and butene, also having similar boiling points, are compared, and the difference is still less between hexane and hexene. Apparently, with the higher members, the effect of the longer chain is more important.

When the molecule contains both the double carbon-carbon bond and the terminal carbonyl oxygen, especially when these are conjugated as in acrolein, the yields are among the highest observed. It is interesting that yield is reduced with crotonaldehyde, in which the carbonyl double bond and the olefin double bond are no longer conjugated. The products are dark and have solubilities similar to those of the saturated aldehydes. In addition, there is often a tendency toward a needle-like structure in the deposit around the slider.

The nitrile group acts similarly to carbonyl oxygen to enhance the yield. Thus, methyl nitrile (acetonitrile) is one of the few methane derivatives which yield polymer, giving 2.4×10^{-5} moles. Acrylonitrile gives 13.2×10^{-5} moles — even more than acrolein.

Acetylene, containing the triple carbon-carbon bond, despite its low boiling point ($-82°C$) and simple structure gives the highest molar yield so far observed, even at room temperature. In this case it is probable that sorption on the palladium is so strong that the acetylene is held readily at a temperature $100°$ above its boiling point. The high chemical reactivity of this substance is well known. The deposits obtained are dark brown and highly insoluble. A C/H atomic ratio of 2.7 was obtained for the acetylene product. Evidently more hydrogen is lost in this conversion than in the case of benzene.

The amine group (NH_2) appears to have a slight but definite depressing action on the polymer yield, especially on the longer chain members.

Apparently hydrogen is not necessary to the production of the polymer. A completely chlorinated unsaturated compound, hexachlorbutadiene, was run with a yield of 0.9×10^{-5} moles, against 1.3×10^{-5} for

butadiene. Triethylamine and its completely fluorinated derivative (pentadecafluorotriethylamine) were run, and about the same volume of deposit was obtained in each case. Hexachlorbenzene also produces polymer.

The yields of polymer obtained from compounds having a cyclic structure are generally higher than those from straight chain compounds having the same number of carbon atoms. Furthermore, the yields are little affected by the degree of saturation or by the presence of hydrocarbon side groups. Thus benzene, toluene, xylene, cyclohexane, cyclohexene, and styrene give practically the same molar yields. The introduction of aldehyde in the side group (benzaldehyde) does little here to raise the yield. Apparently the cyclic structure is the controlling factor. Table IX gives the yields for these compounds.

In reviewing the behavior of the various vapors, it seems clear that factors in the molecular structure which promote polymerization reactions in the bulk material also tend to enhance the yield of frictional polymer. However, it is equally clear that many stable compounds which would not undergo polymerization, except possibly with extreme heat and pressure, readily undergo conversion to frictional polymer. In all cases where polymer was formed, the pronounced lubricating effect already described for benzene was observed. With vapors such as methyl alcohol, carbon tetrachloride and ethylamine which formed no polymer, metallic wear was high, resulting in a deep pit in the plate.

There has been some speculation as to whether polymer formation and activation of contacts may be related. Germer[5] obtained activation on all noble metals including gold and silver, but only with unsaturated cyclic vapors. Polymer is not formed on silver and only in traces on gold, but it is formed with all manner of vapors including saturated and unsaturated cyclic and straight chain compounds. It seems probable that initial adsorption of the vapor is common to both phenomena. From this point on, however, the events take different courses.

TABLE IX — CYCLIC COMPOUNDS — YIELD OF DEPOSIT
(Pd/Pd, 4×10^6 Wipes, 30-gm Force, 120 Cycles/Sec)

Compound	Yield in Moles $\times 10^{-5}$
Benzene	4.9
Toluene	4.9
p-Xylene	5.2
Styrene	4.9
Cyclohexane	5.1
Cyclohexene	4.9
Benzyl Alcohol	4.6
Benzaldehyde	5.5

XI. FRICTIONAL POLYMERIZATION IN THE LIQUID PHASE

In view of the conversion of organic vapors to complex solids on palladium friction surfaces, one naturally wonders what would happen if the sliding were to occur under an organic liquid. The friction devices described for vapor phase studies are not well suited to use under liquids, so another arrangement was devised, in which a palladium wire attached to a vibrating steel reed slides on a stationary wire at right angles to it. The stationary wire is mounted on a phosphor bronze spring which maintains the desired load between the sliding surfaces. The whole device fits into a small glass cylinder, with the wires immersed in about 1 cc of the liquid under study. An electromagnet, positioned outside the tube, drives the steel reed at 120 cycles per second (Fig. 14).

With a 300-micron amplitude under benzene, a copious quantity of brown, insoluble product was produced in four million wipes. This product was identical in all its observable properties to that formed in benzene vapor. Similar results were obtained under butyl alcohol, tetralin, limonene and a light petroleum fraction. The same high degree
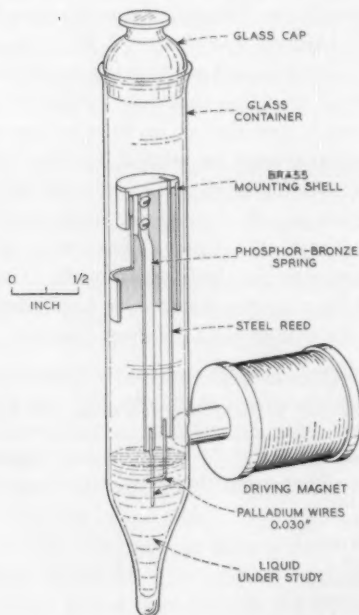

Fig. 14 — Device for generating frictional polymer in liquids.

of lubrication was observed as when the vapors were used — far more than would be expected from the parent liquids.

In all cases, the polymer was held tenaciously to the palladium wires, building up around the point of slide until the mass became great enough to jar off as a sizable flake. There was no indication of any soluble or colloidally dispersed product in the parent liquid. After a 20-hour run in redistilled benzene (16 $\times$ 10^6 wipes), the polymer was filtered out and the liquid was evaporated to recover any soluble matter. None was found. Another portion of the filtered benzene was examined in the infrared spectrograph, with negative results. From these observations, it would appear that the polymer grows to its ultimate insoluble form directly from adsorbed layers on the metal surface rather than through an intermediate soluble or dispersable stage.

As in the case of their vapors, neither methyl alcohol nor carbon tetrachloride produced polymer. In both of these liquids, wear was marked. Since methyl alcohol showed some inhibiting tendencies in the vapor form, it was decided to test this action in the liquid phase, where the concentration ratios of polymer producer and inhibitor can be conveniently and precisely regulated. Mixtures of benzene and methyl alcohol were run, the benzene fraction being increased until polymer definitely formed. It was found that polymer was inhibited completely until a concentration of 95 per cent benzene was reached, confirming the indications obtained with the vapors.

The study of frictional polymerization under liquids is in a preliminary stage at the time of writing. This approach appears to hold considerable promise. The precise control of concentration ratios, the use of reactants not available as vapors and the retention of all reaction products in a small volume of liquid are attractive features. This device also permits studies requiring vacuum-tight, all-glass systems employing the friction mechanism.

XII. FRICTIONAL POLYMER FROM ORGANIC STRUCTURAL MATERIALS

Table X lists a number of tests made on piece parts and materials used in relays or associated with them in wired equipment units. For these tests, the 6-cycle cam-driven device was used and, because of the low yields of polymer frequently encountered, the number of wipes was raised to 14 $\times$ 10^6.

The parts or quantities of the bulk material, equivalent to five relays, were placed in the glass chamber of the friction device with the outlet and inlet tubes left open to simulate roughly the enclosure of a relay in

TABLE X — FRICTIONAL POLYMER FROM VAPORS RELEASED BY ORGANIC STRUCTURAL MATERIALS

(Pd/Pd, 6-Cycle Cam-Driven Device, 14×10^6 Wipes, 170-Micron Stroke)

Piece Part or Material	Temp. °F	Polymer Yield Micrograms
Relay coil with cellulose acetate interleaving	125	34
Relay coil with Mylar interleaving	125	11
Relay coil, wire only (enameled)	125	8
Phenol fiber spoolheads (2C)	120	29
Phenol fiber cards (≠4)	120	22
Phenol fiber cards (≠4) after baking 110°C, 48 hrs.	120	<5
Durez 265 molded relay structures	120	Trace
≠2 Phenolic resin molding compound	120	5
Selectron terminal strip	80	None
Lucite relay contact cover	80	19
Cellulose acetate-butyrate contact cover	80	<5
Neoprene grommets	80	<5
"Scotch" electrical tape, white	80	<5
Wire, cotton, acetate rayon, lacquer (insulation)	80	12
Wire, polyvinyl chloride (insulation)	80	22
Foils		
Laminated Mylar (polyethylene terephthalate)	120	Trace
Koppers polystyrene	80	8
Tenite II, 217A	80	Trace
Benzene saturated air (for comparison)	80	130

an individual can cover. Where the material would normally be subject to heating by relay operation (coil materials, for example), the chamber was heated, with 120 to 125°F being chosen as an average operating temperature.

These results indicate that, when heated, the cellulose acetate interleaving sheet and the phenol fiber spoolheads and cards yield the most polymer. The acetate material is plasticized with diethyl phthalate which distills out in noticeable quantities at 125°C.* When the acetate is replaced by Mylar the polymer yield is reduced to one third, most of which is accounted for by the wire enamel. Baking of the phenol fiber parts markedly reduces the subsequent release of polymer-forming vapors.

Complete control of the polymer by selection and improvement of materials, however, does not offer any great promise because of the large variety of these materials used, not only in the relay but also in associated equipment enclosed with it.

---

* It is interesting that L. H. Germer also has found that the plasticizer in the coil is the chief source of contact activation (private communication).

Attempts were made to adsorb the organic vapors on activated carbon. In practice, the idea did not work out satisfactorily because of the difficulty of locating the adsorbent where it would take up the vapors before they reached the contact. This experience does not rule out the possibility, for special cases, of air being circulated through the adsorbent and blown against the contacts.

XIII. THE PRACTICAL CONTROL OF POLYMER BY SELECTION OF CONTACT METAL

By reference to Table IV it can be seen that, of all the metals that might be considered for use as relay contacts, gold and silver alone might be expected to be free of polymer difficulties. Silver is ruled out for speech transmission because of its tendency to form noise-producing sulphide films. Gold forms polymer, but in quantity so small that it does not interfere in ordinary telephone uses. Under heavy arcing, however, gold erodes badly; also, because of its softness, wear is a problem. In addition, gold is relatively expensive. In all of these respects, except for polymer formation, palladium is an excellent contact metal.

Fortunately, a way out of these difficulties was found by combining gold and palladium structurally so as to obtain the advantages of both metals.[6] The resulting general purpose contact consists of a 1-mil overlay of gold, containing 8 per cent silver, on a 9-mil palladium base. The silver imparts sufficient hardness to the gold to meet the wear requirements. The gold prevents troublesome quantities of polymer from forming when the contacts break little or no current (see Fig. 15). When heavy currents are broken the thin layer of gold will be lost, exposing the erosion-resistant palladium, but build-up of polymer is prevented by the arc. Cost is only slightly increased because only a thin cap of gold is used.

Laboratory and field trials of these gold overlay contacts have proved their superiority in all environments. These trials are discussed in detail in the associated paper by Keefer and Gumley.[2]

XIV. GENERAL DISCUSSION

The cumulative conversion of traces of organic vapors to a visible complex solid when certain metals slide on one another is a startling effect. It seems unlikely that it would have been anticipated from our present knowledge of the physics and chemistry of surfaces, and it has been the primary purpose of this paper to report what is currently known of the phenomenon.

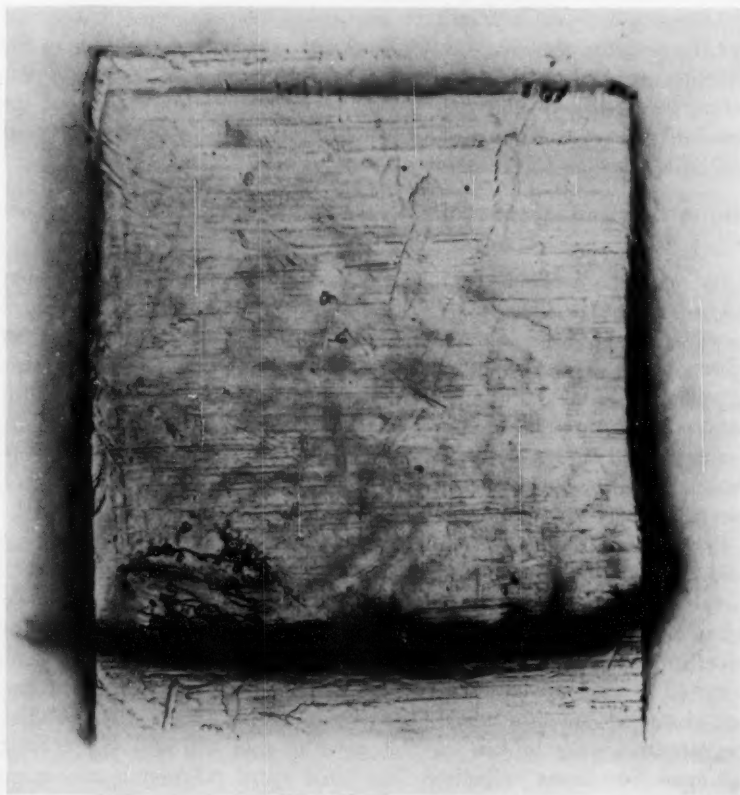Experimental clarification of the mechanism of the frictional poly-

Fig. 15 — Replica of a gold contact operated 10^6 times in the vapors of the organic structural relay parts, showing the small quantity of organic deposit. (Magnified 75 times.)

merization has progressed slowly because of many practical difficulties, not the least of which is the necessity of working on a micro-scale with products not easily characterized chemically. In the absence of an established mechanism, one is left to speculate. Several views have been proposed, none of which is completely without objection. At this point, it might be well to review the observations which need to be explained by any proposed mechanism.

1. When certain metals, notably palladium, are subjected to sliding friction in organic vapors or liquids these are rearranged to form amorphous, colored solids of high molecular weight at the point of slide.

2. This effect is highly sensitive, requiring only trace concentrations of vapors to produce visible quantities of the product.

3. The process is highly efficient, calculations having shown that a monolayer of benzene is converted to the solid product on each palladium surface for each wipe.

4. The conversion occurs with almost every class of organic compound, including many highly stable materials. The presence of hydrogen is not essential; completely halogenated compounds also give a solid product.

5. While the platinum group metals are most active, the effect is not confined to these alone. Molybdenum, tantalum, chromium and gold, for example, also produce solid products. Furthermore, only one of the sliding surfaces need be an active metal. Thus, conversion is obtained when palladium slides on quartz or glass.

6. In all cases, formation of the solid product is accompanied by remarkably effective lubrication.

7. Efforts to inhibit the process by typical catalyst "poisoners" such as carbon monoxide or metallic additions have been unsuccessful. Agents such as benzoquinone and nitric oxide also were without effect. Ammonia, methyl alcohol and carbon tetrachloride in high concentrations prevent or reduce polymer formation by active vapors. However, this suppression of polymer production may simply be the result of monopolizing most of the palladium surface by these strongly adsorbed but non-polymer-producing vapors.

8. Lowering of the temperature has not reduced the yield of polymer in the limited number of cases so studied. On the contrary, in a few instances the yield was increased.

Perhaps the simplest explanation would assume the original molecules to be so strongly sorbed on the metal surface that their structures are put under strain and they are rendered highly reactive. These active molecules then combine among themselves, possibly with simultaneous splitting off of hydrogen or simple hydrocarbons, to produce complex molecular structures. A very thin layer of high molecular weight material thus would coat the surface of any active metal so exposed, the reaction coming to a standstill when the active metal surface was completely covered. The role of sliding friction in this process would be merely to wipe away the reacted product, thus promoting a continuous conversion to polymer at the area of slide. The product wiped away would accumulate at the periphery of the slide area.

This explanation receives some support from studies of the sorption of simple aliphatic and olefinic hydrocarbons on palladium and other metals, in which it has been shown that new compounds are formed.

Typical is the study of chemisorption of ethene on palladium (unpublished work, S. J. Stephens) in which ethane and butane are formed as reaction products. On the other hand, there is no direct evidence to date that such surface reactions have produced the complex, high molecular weight solids encountered as the frictional polymer. Also, where sufficient concentrations of vapor exists, lowering of the temperature has in no case reduced the yield of polymer. To the contrary, certain low-boiling hydrocarbons produce polymer at $-60°C$ but not at room temperature. This negative temperature effect would not be anticipated if chemisorption were a controlling factor.

If the postulated invisible layer of polymer formed directly on the metal surface could be removed without use of friction, the argument for this mechanism would be well established. One attempt was made by the authors, in which a palladium surface under benzene was exposed to supersonic radiation. It was hoped that any polymer film present would be continuously stripped off in this way. Although cavitation was evident at the palladium surface, no polymer appeared in the benzene, infrared examination of which showed no change.

The failure to prevent production of the deposits by recognized catalyst poisons would also seem to argue against the mechanism of direct polymerization.

Another objection to this explanation is the assumption that the wiping action re-exposes the catalytically active metal surface. If this is so, one should expect considerable wear to attend the production of the polymer. However, it should be recalled that a high degree of lubrication is one of the most striking and consistent effects associated with its production. It seems quite probable that a large part of the slide is not that of metal on metal, but that a film of some sort is interposed. Nevertheless, calculations have shown that about a monolayer of benzene is transformed to polymer per wipe on each surface. It would be difficult to reconcile this high efficiency of polymer production with the high efficiency of lubrication if metal-to-metal wipe were a necessary condition.

This objection might be met by assuming slide to take place principally between chemisorbed "primary" organic layers, through which the influence of the active metal is still exerted on superimposed layers. On wiping, shear might then take place at or near the primary layer, leaving it intact but highly reactive chemically. In such a mechanism, the role of friction becomes more important. This suggests the possibility that part of the work of sliding friction may go into raising the chemical activity of the adsorbed organic layers, with resulting formation of the

high molecular weight products. Such an explanation would admittedly lack the simplicity of the direct polymerization idea, since it would involve two or more steps. Certainly the sorption of the organic molecules on specific metal surfaces would be the first step. Work then performed on these layers by friction would convert the sorbed layers to polymer in one or more additional steps.

Just how frictional work would bring about polymerization is a highly speculative question. While friction is a well known phenomenon from an engineering standpoint, the changes taking place on a molecular scale when one surface slides over another are certainly not fully understood.

Where friction is involved, one immediately thinks of heat and, in the present case, of the possibility that temperatures might be reached at which organic molecules would be broken down into active fragments with ultimate recombination into more complex structures. However, the slide velocities at which polymer forms are such that the temperature rise for the clean metals would be less than 50°C on the basis of Bowden's measurements,[7] and, in this case, the temperature would be even lower because of the lubricating effect of the polymer. A run was made on the cam-driven machine using benzene-saturated air for 4×10^6 wipes at 22 wipes per minute. This corresponds to a slide velocity of only 0.006 cm/sec. A full yield of polymer (> 40 micrograms) was obtained, giving further evidence that frictional heat is not necessary. Already mentioned are the polymer yields obtained at $-60°C$. Experiments were also performed in which the operating temperature of the machine was raised to 150°C with no effect on the yield of the benzene polymer. Unsuccessful attempts were also made to form similar polymers without friction on platinum and palladium surfaces subject to cycles of heating and cooling. Thus it does not appear that frictional heat, at least as we understand it, produces the polymer.

The action of sliding one metal surface over another is usually thought of in terms of gross mechanical effects, such as seizure and galling, which of course are very important in lubrication engineering. These effects are believed to be the cumulative result of shear motion and the strong molecular attractions which occur when the surface structures are forced into sufficiently intimate contact. When chemisorbed organic films cover the sliding metal surfaces it seems quite possible that the interplay of attractive forces with straining and deformation of the molecules may supply the activation energies necessary to bring about drastic chemical rearrangement. Such a mechanism would place emphasis on sliding friction as the primary source of the polymerization, although

strong bonding of the organic molecules to the metal surface would be a requisite. Associated with this "frictional working" of the adsorbed layer would probably be a profound disturbance of the valence electrons. The possibility of free radicals existing in the friction product was suggested by D. Tabor[8] and was experimentally explored by the writers. The deposit generated from benzene vapor on palladium surfaces was allowed to fall into a capsule containing acrylonitrile, through which a Tyndall beam was passed. It was assumed that any free radicals in the deposit would initiate mass polymerization in the acrylonitrile with the development of increasing turbidity. This failed to occur. Liquid acrylonitrile was also subjected to palladium-to-palladium slide in the "capsule" device already described. There was no evidence of any polymerization beyond the actual friction area. These experiments do not rule out the possibility that something akin to free radicals may exist in a "captive state" on the metal surfaces, held there by adsorptive forces.

XV. ACKNOWLEDGMENTS

REFERENCES

1. Hermance, H. W. and Egan, T. F., *Proceedings of 1956 Electronic Components Symposium*, Engineering Publishers, N. Y., 1957, pp. 85–93, also Communications and Electronics, **34,** Jan., 1958, pp. 756–763.
2. Keefer, H. J. and Gumley, R. G., this issue, pp. 777–814.
3. Wilson, R. W., The Contact Resistance and Mechanical Properties of Surface Films on Metals, Proc. Phys. Soc. (London), **68B,** September, 1955, pp. 625–641.
4. Barns, R. L. (Bell Tel. Labs.), unpublished work.
5. Germer, L. H. and Smith, J. L., Activation of Electrical Contacts by Organic Vapors, B.S.T.J., **36,** May, 1957, pp. 769–812.
6. Egan, T. F., U. S. Patent 2,812,406.
7. Bowden, F. P. and Tabor, D., *Friction and Lubrication of Solids*, Oxford Univ. Press, N. Y., 1954.
8. Tabor, D., private communication.

Relay Contact Behavior Under Non-Eroding Circuit Conditions

By H. J. KEEFER and R. H. GUMLEY

(Manuscript received July 3, 1957)

When palladium or platinum-family metal contacts are operated in an environment which includes organic vapors, the performance of the contacts may be seriously degraded. If the contacts open and close currents, the contact erosion may be increased enormously. On the other hand, if the contacts merely prepare the circuit and do not directly open or close the current flow, no erosion will take place. The contacts may then fail to conduct the circuit current satisfactorily, because of the formation of an insulating polymer-like film. This latter type of failure has been the subject of extensive studies at Bell Telephone Laboratories for the past several years and is the subject of this article. These studies have yielded a new understanding of the contact problem and resulted in certain changes in relay contacts to improve their performance under service conditions.

I. HISTORICAL

The effect of organic vapors in degrading the performance of platinum metal relay contacts has long been recognized. In a paper published in 1927 by E. A. Watson,[1] it was shown that carbon readily deposits as a closely adherent film on the contact surfaces, increasing the tendency of the contacts to arc and greatly accelerating the electrical erosion. An early investigation into the effect of organic vapors in enhancing contact erosion is described by A. Brooks in a discussion of a paper by J. C. Chaston.[2]

Prior to the introduction of the No. 1 crossbar system in the late 1930's, the effect of organic vapors on relay contacts was not of great concern in Bell System equipments. Dust was considered to be the main cause of contact failure, the requirements for contact reliability were not as exacting and automatic means for detection of contact failure were not as commonly employed. Except for some unusual effects that

occurred at the time some laboratory rooms were painted, the effect of organic vapors on relay contacts attracted little interest.

Toward the end of the 1930's, the No. 1 crossbar local switching system was developed. This system and all subsequent crossbar systems were of the common control type, using markers that put increasing demands on the operating speed, life and reliability of the relays. The relay contacts had to operate faster and more often, switch larger currents and last longer. With the No. 1 crossbar system came the U-type relay using cellulose-acetate-filled coils that increased the quantity of organic vapors. However, the effect of these vapors on the contact performance was not immediately apparent.

In 1943, an attempt was made in the Switching Systems laboratory to increase contact reliability of U-type relays by enclosing the relays in sealed containers for dust exclusion. It was found, however, that the erosion of the relay contacts was increased 20 times or more, because of the increased concentration of organic vapors. A series of tests was initiated to explore this effect, and a summary of some of the effects has been published.[3] Simultaneously and independently, the Telegraph Development Department laboratory was finding that new organic materials incorporated into the 255-type telegraph relay had greatly accelerated the erosion of the relay contacts. In 1948, contact erosion tests on the prototype of the wire-spring relay indicated the need for a contact cover that did not enclose the coil. This contact cover is a feature of all wire-spring relays now used in switching systems. Since the war, the effect of organic vapors in accelerating contact erosion has been intensively studied by L. H. Germer and his associates at the Laboratories.[4]

During this time, it was known that the deposits from the organic vapors could cause some contact resistance, but this was considered to be a much less serious fault than the erosion effect. In 1948, an analysis by means of the replica technique[5] of failing contacts in service, disclosed that a polymer-like* type of deposit was formed on palladium contacts, particularly on those contacts that were not subject to electrical erosion. A replica taken from an AF relay contact showing the polymer formation is illustrated in Fig. 1. The highly insulating properties of this material indicated a previously unknown cause of contact failure, but the importance of the discovery could not be estimated until large-

---

* The mechanism of generation of this insulating material is not completely understood. For the purposes of this paper it is necessary to distinguish between this insulating material and other organic deposits such as the carbonaceous deposits that result when contacts arc in the presence of organic vapors. Since this insulating material has many of the attributes of a polymer it will be referred to as such.

Fig. 1 — Replica of an AF relay twin palladium contact, showing polymer formation at the contact area. (Magnification 40 times.)

scale relay operating tests were made. Nevertheless, an investigation into the physics and chemistry of this organic deposit was immediately started. Results are contained in a companion article.[6]

In 1951, tests were made in the laboratory on a large number of U-type relays in various types of dust enclosures. Results showed that the failures caused by polymer were more numerous than the dust failures.[7] Nevertheless, it was felt that the U-type relay had fairly good contact reliability in telephone service and, since dust appeared to be the main trouble under service conditions, no change involving increased costs could be justified to reduce the polymer contamination of the contacts.

In 1951, it was found that certain contacts carrying talking currents had become microphonic and under some conditions noisy in service. Investigation showed the noise to be due to polymer and carbonaceous deposits on the contacts which, when vibrated by the operation of adjacent relays, generated the noise.

In 1953, large-scale contact-reliability tests on pre-production wire-spring relays were started. In these tests, the wire-spring relay was found to have a higher failure rate from polymer contamination than did the U-type relay. To avoid risk of potential service difficulties, a

1-mil 22-karat gold overlay was specified over the 9-mil palladium contacts on wire-spring relays, because only a trace of the polymer forms on gold. For economy, the standard wire-spring relay design uses the gold overlay only on the moving contacts. Laboratory tests have shown that this compromise is satisfactory and that the relay, so equipped, will outperform the U-type relay equipped with palladium contacts, from an open-contact standpoint.

In 1955, open-contact failures were experienced in service with palladium contacts of digit-absorbing selectors of the step-by-step system. These failures were particularly serious, since they were of the type that deny service to the customer. The contacts were changed to Western Electric No. 1 metal,* since this contact metal was known to give substantially improved performance and could be made available quickly. No further difficulty has been experienced from these contacts.

The laboratory studies to date have indicated that gold, silver or their alloys are the best solution to the polymer problem. The use of an alloy of high gold content is preferred because of the undesirable sulphiding characteristics of silver. The gold overlay used on the wire-spring relays is 22-karat gold, the 8 per cent silver content being solely for hardening purposes. The gold-overlay-on-palladium contact is preferred to a solid gold-alloy contact such as Western Electric No. 1 metal for general use on relays because No. 1 metal is not as good from an erosion standpoint and is more costly than palladium.

Polymer failures are not likely to occur on contacts which erode, since the arcing burns away the polymer. However, the gold overlay is applied to all wire-spring relays because it is impractical to know in advance the use to which the contacts of each relay code will be put. About 75 per cent of all contacts in switching systems do not erode, and these contacts will benefit by the gold overlay, which need only be thick enough to provide for the expected mechanical wear. The other 25 per cent of the contacts will erode and, for these, the gold serves no useful purpose. These latter contacts will erode the 1-mil overlay of gold fairly quickly and then obtain their needed erosion life from the underlying 9-mil-thick palladium metal.

II. GENERAL THEORY

When palladium or platinum-family metal contacts are operated in an environment containing organic vapors, a polymer-like substance is formed on the contact surfaces. The polymer forms only on operating,

---

* An alloy of 69 per cent gold, 25 per cent silver, 6 per cent platinum by weight.

sliding or vibrating contacts. The polymer generation does not depend on the contact current, since it is also produced on unwired contacts. A detailed physicochemical description of the polymer generation and properties is contained in a companion article.[6]

The polymer accumulates in compacted clumps around the actual contact area or in dust on the sides of the contacts. Enough will be generated in 1000 operations of a relay contact to be visible under a microscope. In 100,000 operations, the material is easily visible with a 10-power glass. After several hundred thousand operations, the compactions are large enough to cause contact failure in 50-volt circuits by dusting off and falling into the contacting area. After possibly a half million operations, the polymer dusts away from the contacts at the same rate as it is produced, so that the contact failures then result at a maximum rate. Failures from polymer may also occur because of a shift in the actual contact area caused by a slight change in the contact spring position or a wearing down of a minute metallic roughness in the contact surface. The many factors that affect the failure rate are outlined in Section IV.

The effect of the polymer on actual operating relay contacts has been studied intensively only in 50-volt circuits, since this voltage is most commonly used in telephone switching practice. In such circuits, a contact, having failed, will clear itself, on the average, in about five relay operations. Thereafter, its probability of failure is not much greater than that of other contacts in the same test. This low "persistency" of failure is characteristic of the polymer, whereas fibrous dust failures tend to be much more persistent. Although lower circuit voltages have not been fully explored, it is expected that the failure rates and the persistency of failure will increase as the voltage is decreased.

The polymer is a good insulator and has the characteristics described by Holm[8] for thick highly insulating films. When voltages below the dielectric breakdown value are applied, the resistance of the film is many megohms and the currents are zero or a few microamperes. If the voltage is raised sufficiently, dielectric breakdown occurs, followed by the formation of a metallic filament pulled out of the contact surfaces (coherer effect), and the contact voltage then drops to a value corresponding to the melting-point voltage of palladium, about 0.5 volt. The voltage required to produce such breakdown through the polymer particles has been found to be as high as 270 volts.

Often the polymer is mixed with carbonization or erosion products which lower the film resistance. The polymer film may be carbonized by electrical arcing even on "non-eroding" contacts since the circuit voltage,

when applied to the insulating film, may arc through the film, thereby decomposing it by heat. Also, erosion products form when the relays are operated under conditions of low vapor concentrations and high pulsing rates. The resulting effect is the same as for lightly arcing contacts, in that film resistances of a few ohms are commonly found and high resistances are less frequently obtained.

When the organic deposit is highly carbonized, the contact resistance is usually low and the current is proportional to the voltage. However, if the current is raised sufficiently, thermal breakdown will occur, with an abrupt drop in resistance to a new lower value. The process can be repeated until the melting-point voltage is reached.

III. FAILURE RATES AND UNITS

A 10,000-line No. 5 crossbar central office handles about 50,000 calls per day. Each call requires the operation of about 1000 relays, each relay having an average of seven contacts. Consequently, such a central office has about 100×10^9 relay contact operations in one year.

Maintenance data indicate that such an office, using U-type relays, has about 100 "found" open contacts per year, or, using the above contact operations, about one found open contact per 10^9 contact operations. It is estimated that there are at least 12 times as many failures as this, but these other failures are not persistent enough to be found by the maintenance forces and appear only as transient failures.

The laboratory tests to be described are more efficient at locating troubles than are central office trouble detecting routines. It is estimated that half of the indicated troubles in these laboratory tests are found. Consequently, in these laboratory tests a found trouble rate of about six opens per 10^9 operations would be equivalent to central office experience, but failure rates much higher than this would be a cause for concern.

In the tests to be described, the failure rates are listed on the basis of "opens per 10^9 contact operations". Since each test usually involves several hundred contacts operating several million times each, it would be more descriptive to use "opens per thousand contacts per million operations" as a unit. However, the simpler unit is used for brevity.

In all cases, only the initial failure of a given contact in a series of consecutive contact failures is counted in the failure rate. The successive failures that occur after an initial failure and until the contact clears itself are counted as the "persistency" of failure. The total failures are therefore the failure rate times the average persistency.

IV. FACTORS AFFECTING CONTACT PERFORMANCE

The factors that affect or determine the contact performance of any pair of relay contacts can be classified into four categories: (1) the contacts, (2) the environment, (3) the dynamics of the contacts and (4) the circuit connected to the contacts.

Insofar as the contacts themselves are concerned, it is necessary to consider their size, shape, alignment, smoothness, hardness and composition, as well as the presence of any erosion products, films or particulate matter on the contact surfaces. The contact environment includes the atmosphere, which may be dusty, and any additional vapor or gas, including humidity, that would form films or deposit other foreign material on the contact surfaces. The contact dynamics include the velocity, impact, slide, chatter, force and rate of operation of the contacts. The circuit conditions involve the voltage, current, load inductance and load capacitance and wire capacitance, and whether the contacts close and open with current flowing.

The effect of many of these variables on contact performance in the presence of organic vapors will be described in later sections.

V. METHODS OF TESTING

Many of the variables listed above were explored by direct contact-reliability tests on actual relays. Three types of test setups were used. For convenience, these are called the Open Contact Test Machine (OCTM), the Open Contact Test Unit (OCTU) and the Open Contact Test Frame (OCTF). The OCTM is used for exploratory tests of several months' duration. Single contacts are used to increase the failure rate and only one parameter is varied in each test of 90 contacts. About 40 such tests have been made. The OCTF was used for a direct comparison test of several years' duration. A total of 2000 single and twin contacts on four different relay types were tested simultaneously under a fixed set of circuit and environmental conditions. The OCTU simulates actual field conditions and is operated for a one-year period, using 540 twin contacts under the equipment conditions used in the telephone plant. Ten such tests have been made. Simplified schematics of the OCTM and OCTU are shown in the Appendix (Figs. 18 and 19).

In these tests, the contacts are generally operated and released without contact current so that no contact erosion occurs. However, in some tests the test contacts are allowed to charge the wire capacitance of the succeeding contacts in the series contact chains. This type of contact closure is referred to as "cable charge." The tests in which cable charging
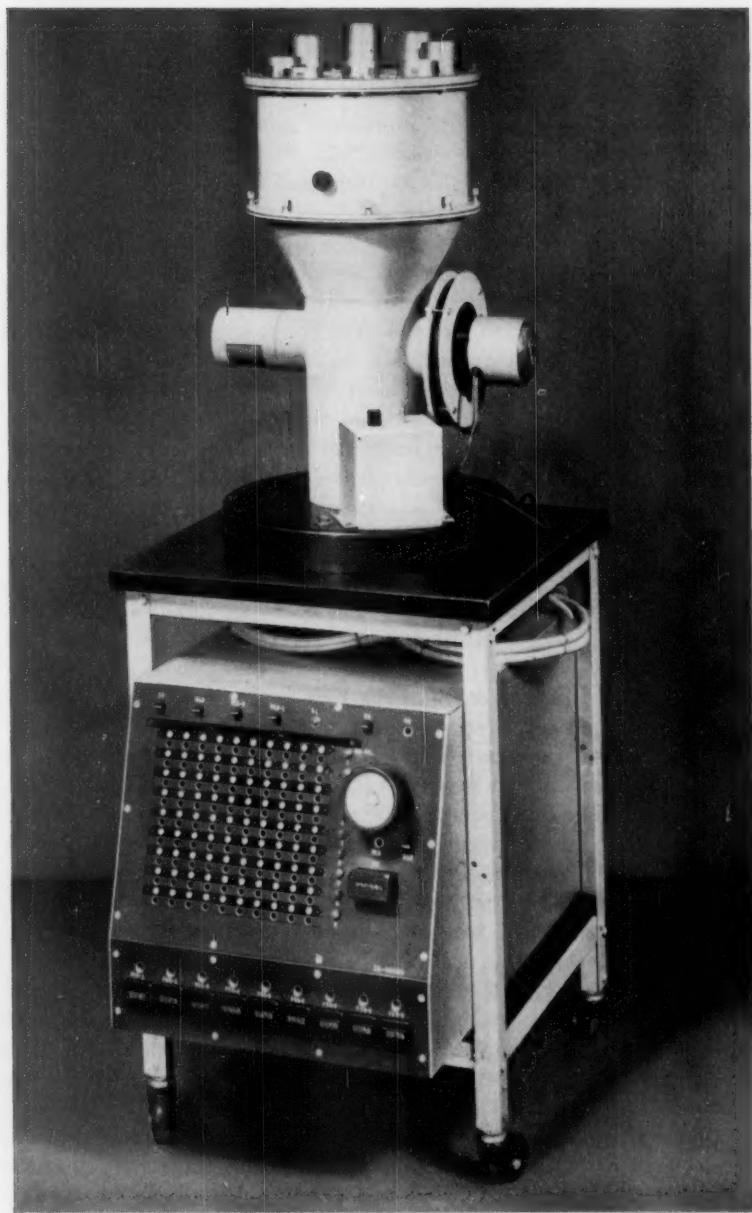
Fig. 2(a) — Open Contact Test Machine (OCTM) is used for exploratory tests of short duration. Single contacts are used instead of twins to increase the failure rate, with usually only one parameter being varied for each test.
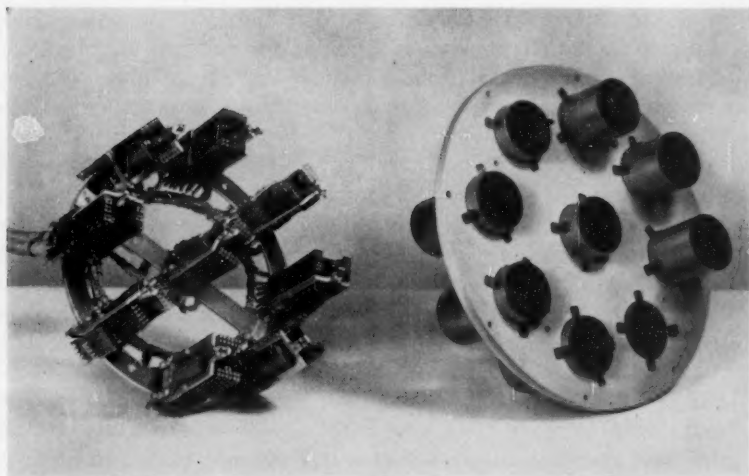
Fig. 2(b) — Close-up of nine test relays and associated streamlining cylinders which are mounted in the upper turret of the OCTM. The streamlining cylinders were removed and a blank plate substituted for these sealed-chamber tests.

by the test contacts is not permitted are referred to as "no cable charge." The test sequence is generally as follows: (1) the test relays are operated, (2) a 50-volt checking circuit is applied through the contacts arranged in series chains, (3) the checking circuit is removed, (4) the relays are released, (5) the cycle is repeated. The cycle is controlled by a mechanical interrupter. If a contact fails, the circuit stops in the failed position until this failure is located, and the circuit is released by the attendant. Unless otherwise indicated, each test is started with new relays and the contacts are replica-cleaned to insure freedom from initial contamination.

The Open Contact Test Machine (OCTM), Fig. 2, consists of 90 normally-open test contacts arranged in series chains of ten contacts per chain. These test contacts are on nine relays that are housed in the upper turret of the test machine. The circuit operation is arranged to stop and hold in the failed position when a contact or series chain of ten contacts produces a resistance greater than 1000 ohms. The checking circuit placed across the series chain of test contacts provides approximately 0.020 amperes at 50 volts. The time the checking circuit remains across the contacts can be varied from 30 to 125 milliseconds.

The OCTM was originally designed with the objective of using it as a dust meter to measure the effect of dust in causing opens on relay con-

tacts. Dusty air can be drawn downward through the streamlining cylinders past the relay contacts while the relays are being operated. For testing organic deposit contamination, the streamlining cylinders are removed and the upper surface is sealed with a tight-fitting plate to obtain a sealed chamber.

The 90 contacts on the nine test relays are wired individually to 90 jacks located on the lower front of the test machine. If a contact fails to close, the machine stops and the failure is indicated by an alarm. An attendant then applies a shunt counting circuit across the failing contact to allow the operation to proceed. The counting circuit counts the number of consecutive contact failures, as a measure of the failure persistency. When the contact open is cleared by these operations, the shunt circuit is disconnected automatically.

The usual rate of operation with the octm is at one operation per second. Prior to the start of a new test, the relays are carefully cleaned of dust and fibers by vacuum cleaning and the contacts are cleaned by the replica method. Furthermore, the chamber is never opened once the test is started and the contact troubles are cleared only by the successive operations of the contacts themselves.

The Open Contact Test Unit (octu), Fig. 3, consists of 90 test relays, each having three make and three break contacts arranged in series chains of 30 contacts per chain. The relays are enclosed in the covers which are standard for the system in which the relays are used. Thus, in testing wire-spring relays, no frame covers are used, but each relay has its own molded-plastic contact cover. The circuit operation is arranged to stop and hold in the failed position when a contact or series chain of 30 contacts produces a resistance greater than 1000 ohms. The checking circuit, placed across the series chain of test contacts for 30 milliseconds, consists of approximately 0.080 ampere at 50 volts.

If a contact fails to close, the test stops and the failure is indicated by an alarm. An attendant then locates the failing contact by use of an ohmmeter. Failing contacts are cleaned by the replica method. Failure persistency data are, therefore, not obtained in the octu.

The usual method of testing with the octu is testing of twin contacts at 0.1, 1.0, and 10 operations per second. Prior to the start of a new test all contacts are cleaned by vacuuming and then by the replica method.

The Open Contact Test Frame (octf), Fig. 4, consists of 200 test relays, each having five make and five break contacts arranged in series chains of ten contacts. The octf is made up of two bays, one housing the relays under test, the other housing both the control relays and the jack field used to obtain access to the failing contact. The test bay is
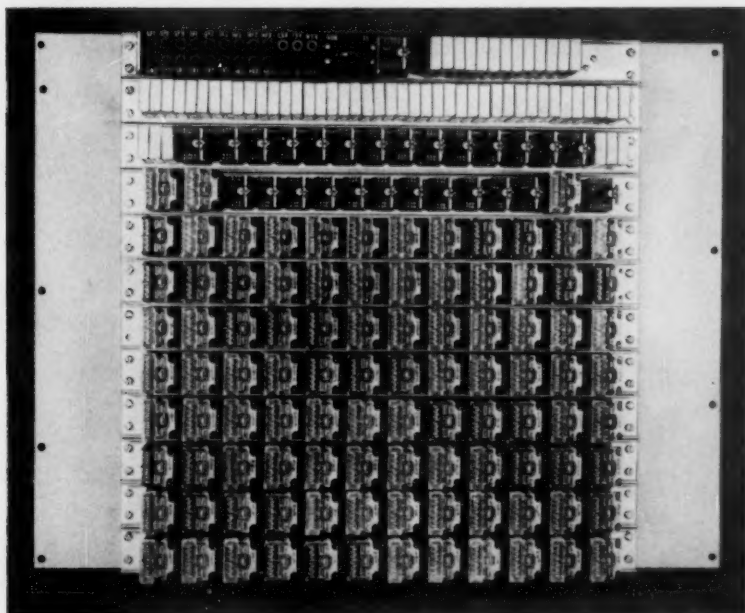
Fig. 3 — Open Contact Test Unit (octu) simulates actual field conditions in that twin contacts are operated under equipment conditions used in the telephone plant. Tests are usually operated for more than a year, with contact parameters indicated by exploratory tests in the octm's.

sealed and gasketed, with the relays molded into the mounting plates to preserve a dust-tight enclosure. Controlled dusty or filtered air can be admitted by means of blowers and ducts, but this facility was not used in these tests. Four types of general purpose relays, two U-type designs (U and UB) and two wire-spring designs (AF, and M24*), were tested simultaneously, using both single and twin palladium contacts. The circuit operation is similar to that of the octu. However, the location and clearing of a failing contact is like that of the octm, where failure persistency is obtained. One test was made which included 29 months of operation and the testing of a total of 2000 contacts.

VI. TEST RESULTS

The tests to be described were mainly exploratory in nature, and were designed to indicate trends rather than absolute values. As might be expected in making open-contact failure comparisons, the variability of

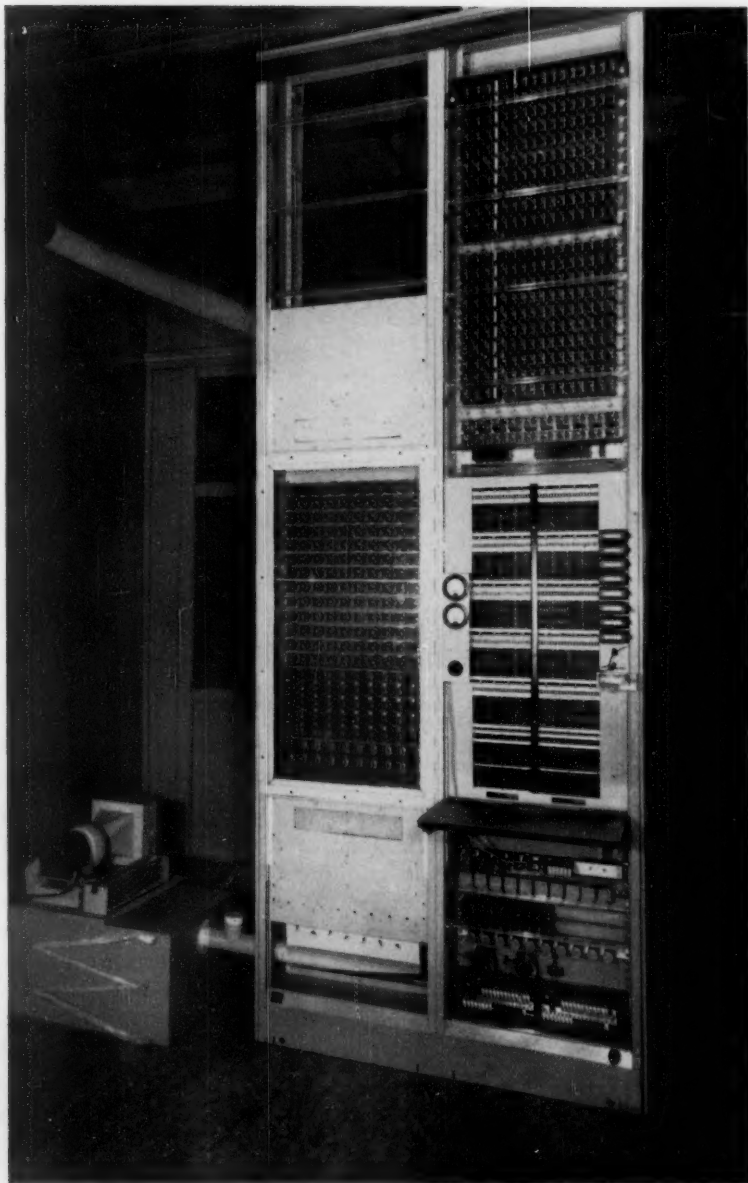* An early wire-spring relay design that was not used in final production.

Fig. 4 — Open Contact Test Frame (octf) was used for a direct comparison test of several years' duration. Single and twin contacts on four different relay types were tested simultaneously under a fixed set of circuit and environmental conditions.

the data is large, so that differences in performance of less than 2:1 may not be significant.

6.1 *Effect of Kind of Contact Metal*

Palladium contacts are relatively poor performers when subjected to polymer contamination by organic vapors. On the other hand, gold, silver and gold-silver alloys are relatively trouble-free from a polymer standpoint. In one series of tests, these contact metals were tested in the form of heavy-size single contacts on U-type relays in the OCTM units. The alloys used were No. 1 metal (69 Au, 25 Ag, 6 Pt), No. 3 metal (70 Au, 30 Ag), 22-karat gold (92 Au, 8 Ag), and R156 metal (60 Pd, 40 Ag). The results of these tests are shown in Table I.

TABLE I — FAILURE RATES WITH CERTAIN CONTACT METALS SINGLE CONTACTS ON U RELAYS

Contact Metal	Cable Charging on Closure	Relay Operations in Millions	Operations to First Open in Millions	Opens/10^9 Contact Operations	Average Persistency
R156/R156	Yes	1.82	0.07	565	3
Pd/Pd	Yes	2.50	0.31	218	6
No. 1/No. 1	Yes	10.00	0.79	15	2
Pd/Ag	No	1.83	1.03	6	2
No. 3/No. 3	Yes	10.55	—	0	—
22K Au/22K Au	Yes	10.00	—	0	—

A similar series of tests was run on AF-type (wire-spring) relays. In this series of tests, the gold-silver alloys were mostly overlays on palladium. Also tested were certain combinations of these gold-silver alloys mating with palladium contacts. The results of these tests are shown in Table II.

The large improvement obtained with palladium mating with 22-

TABLE II — FAILURE RATES WITH CERTAIN CONTACT METALS SINGLE CONTACTS ON AF RELAYS

Contact Metal	Cable Charging on Closure	Relay Operations in Millions	Operations to First Open in Millions	Opens/10^9 Contact Operations	Average Persistency
Pd/Pd	Yes	1.55	0.10	4250	4
Pd/No. 1	Yes	0.78	0.51	358	16
Pd/22K Au	No	6.23	1.45	271	15
Pd/No. 3	Yes	0.78	0.39	32	10
No. 1/No. 1	Yes	5.08	2.32	4	2
No. 3/No. 3	Yes	10.00	9.30	1	5
22K Au/22K Au	No	3.00	—	0	—
Ag/Ag	Yes	7.28	—	0	—

karat gold indicated that this combination might perform very well as twin contacts. This is the combination that was finally used on wire-spring relays. The final test using twin contacts is described in a later section.

6.2 *Effect of Contact Shape*

A series of tests was made of various shapes of palladium contacts on U-type relays to investigate the possibility of obtaining improved performance over the performance of the present heavy palladium (large-area) contacts. It was found that no worthwhile reduction in polymer failures could be obtained by changing the contact shape.

These tests were conducted in the octm units. The contact shapes tried, some in combinations, are shown in Fig. 5. In all cases, the mating contacts were assembled on the springs so as to form crossed-bar contacts.

The results of these tests, all with palladium contacts, are summarized in Table III.

It is evident that the contact shape that tends to produce the maximum contacting area produces the smallest failure rate from polymer.
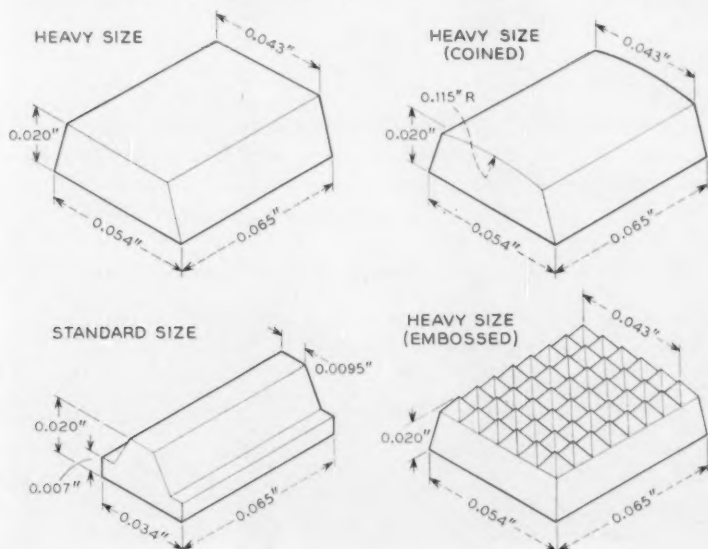

Fig. 5 — Various shapes of palladium contacts used on U-type relay tests in the octm's.

Table III — Failure Rates with Certain Contact Shapes
Single Palladium Contacts on U Relays

Contact Form	Cable Charging on Closure	Relay Operations in Millions	Operations to First Open in Millions	Opens/10⁹ Contact Operations	Average Persistency
Coined/Coined	Yes	0.60	0.20	1196	5
Coined/Hvy	Yes	0.34	0.17	391	5
Emb/Hvy	Yes	1.06	0.13	385	4
Std/Std	Yes	1.27	0.27	306	3
Hvy/Hvy	Yes	2.50	0.31	218	6

Similar results were found in earlier contact tests[9] made at the Laboratories, in which the contact contaminants were limited to dust particles less than 25 microns in diameter. On the other hand, in these earlier tests, when the contaminating dust was changed to larger-particle lint and paper fibers, the opposite trend for the effect of contact shape was observed.

An explanation for the observed effect of the contact shape on the failure rate from polymer is deferred to the next section. It should be noted that, in addition to changing the contact shape, the coining and embossing operations work-harden the contact surfaces and improve the microscopic smoothness. These factors also affect the failure rate.

6.3 *Effect of Contact Alignment*

Typical distributions of open-contact failures on palladium contacts of U and AF relays in octm tests are shown in Fig. 6. This graph shows that some contacts are relatively prone to failure while others, supposedly identical, never fail. Attempts have been made to explain why contacts differ so greatly in their performance. It is speculated that the difference between such contacts is partly one of alignment of the contacting surfaces, and that polymer-caused open-contact failures on palladium contacts would be minimized if the contacts could be made flat and be aligned so as to tend to produce a maximum contacting area.

A study of contact replicas taken from U-relay contacts at the completion of a test supports this speculation. Contacts which have a history of contact failures produce replicas like those of Fig. 7(a), while contacts that have not failed produce replicas like Fig. 7(b). It appears that, with good alignment, the polymer films remain thin and the true conducting spots may be spread over quite an area. These conducting spots are very small, being on the order of 10 to 20 microns in diameter. If one such conducting spot wears down or becomes insulated by a fine polymer dust, electrical contact may be made via the other conduct-
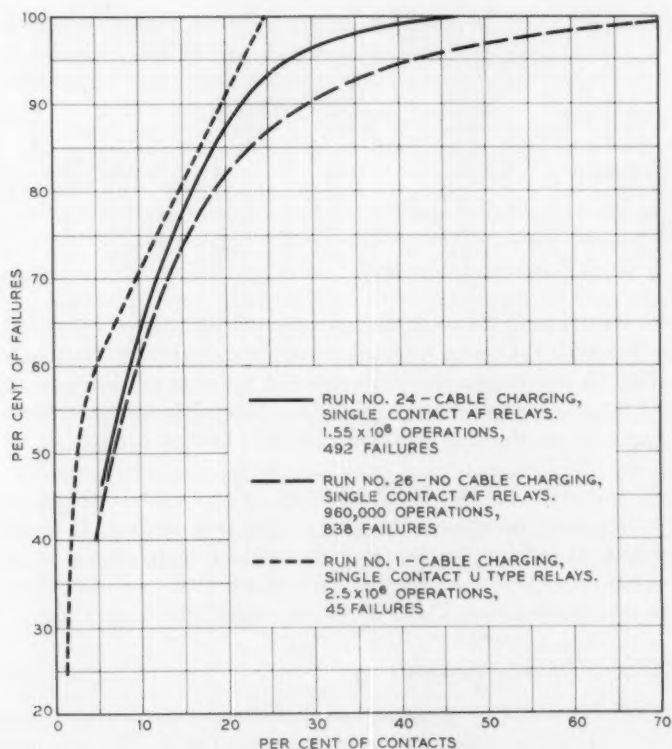
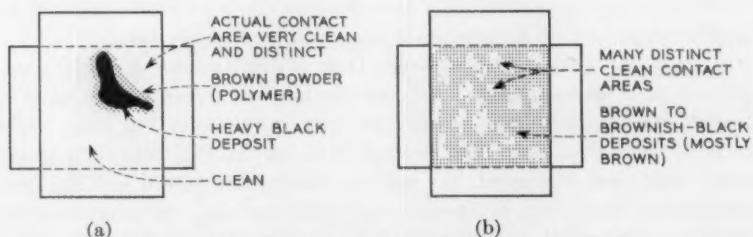Fig. 6 — Distribution of open-contact failures on palladium contacts of U and AF relays in ocTM tests.

Fig. 7 — Sketches of replicas of single contacts of U-type relays: (a) poor alignment — produces opens (b) good alignment — does not produce opens.

ing spots, and the contact will not fail. With poor alignment, the contacts form compacted aggregates of polymer surrounding a very limited contact area. If the true conducting spots wear down, another electrical contact cannot be made in the compacted ring of polymer, and the contact fails. Also, the compacted polymer may break off and cause contact failure by falling into the contact area. Contacts which have been coined or embossed to present only a small contacting area behave like misaligned flat contacts, and therefore have a relatively high failure rate, as indicated in the results of the previous section.

6.4 *Effect of Operating Rate*

The operating rate is a major factor in the open-contact failure rate. When the operating rate is low, not enough polymer forms and the

TABLE IV — FAILURE RATES WITH THREE OPERATING RATES
TWIN PALLADIUM CONTACTS ON AF RELAYS

Number of Test Contacts	Operations per Second	Relay Operations in Millions	Total Opens	Operations to First Open in Millions	Opens per 1000 Contacts/Year	Opens per 10⁸ Contact Operations
120	0.1	0.39	4	0.15	33	86
120	1.0	3.88	48	0.59	400	103
120	10.0	38.80	13	1.59	108	3

failure rate, per unit of time, is low. If the operating rate is high, the composition of the organic deposit apparently changes and the failure rates are also found to be low. However, at intermediate operating rates such as one operation per second, sufficient polymer with highly insulating properties is formed to produce a high failure rate per unit of time. The data shown in Table IV were derived from an OCTU, using AF-type relays having twin palladium contacts and individual molded-plastic contact covers but no frame covers. The test was operated for one year.

Since this test showed that failure rates per unit of time are greatest with a relay operating rate of about one operation per second, and since the common control circuits of crossbar systems use relays at about this operating rate, most of the other tests were conducted using a rate of one operation per second. Several attempts were made to accelerate the tests by using higher operating rates and by applying a correspondingly increased vapor concentration with the use of limonene, as had been done in the study of the effects of organic vapor on contact erosion.

The increased vapor concentration, however, apparently affected the physical makeup of the polymer, so that it was kept more fluid and did not produce the failure rates obtained at the lower operating rates. These accelerated tests were therefore discontinued.

6.5 *Effect of Operational History*

The failure rate is significantly affected by the previous operational history of the contacts. The failure rates are low until several hundred thousand operations have produced enough polymer to cause open-contact failures. Thereafter, the maximum failure rates are encountered, sometimes decreasing gradually with operations. As will be shown in other tests to be described, the decrease in failure rate may result from somewhat better alignment or fit of the contact surfaces, resulting from mechanical and electrical erosion of the contact surfaces. It is not due primarily to a general drying out of the volatile components of the relay structures, although this may be a secondary cause of the observed effect. The following data were derived from an OCTU with AF-type relays using twin palladium contacts and individual molded-plastic contact covers but no frame covers. This OCTU had had a year's operation (the test described in Section 6.4), with the relays being divided into three groups, each having a different rate of operation, as shown. Before the start of the new test, all contacts were replica-cleaned. The test was then restarted with all relays operating at the uniform rate of one operation per second, with results shown in Table V.

The performance of the contacts that had the history of 38,800,000 previous operations is remarkably good. The improvement does not seem to be due to a reduction in the ability of the contacts to produce polymer, because, at the end of the test, a replica examination failed to show any difference in the amount or kind of organic deposit formed on the contacts of the three groups. In the succeeding sections several tests are described which explore this effect further.

TABLE V — FAILURE RATES WITH DIFFERENT OPERATIONAL HISTORIES
TWIN PALLADIUM CONTACTS ON AF RELAYS

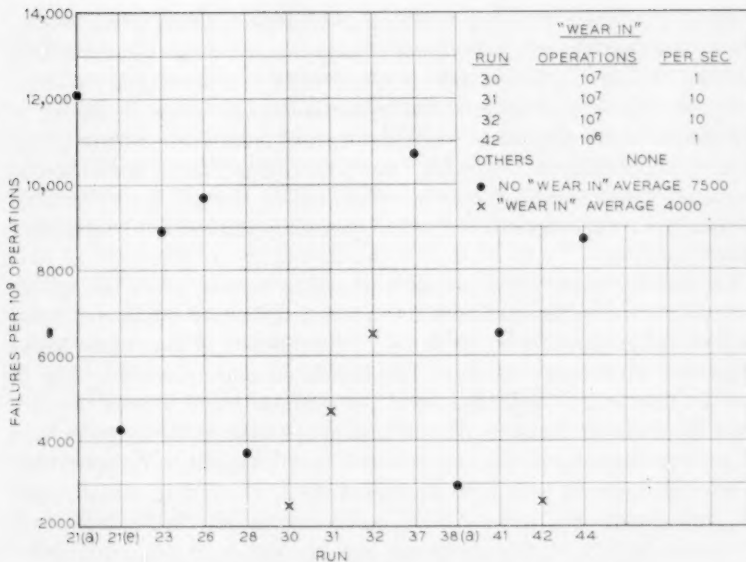| Previous History		New Test (After Cleaning)			
Operations per Second	Relay Operations in Millions	Operations per Second	Relay Operations in Millions	Number of Opens	Opens Per 10^9 Contact Operations
10.0	38.8	1.0	2.0	0	0
1.0	3.88	1.0	2.0	10	27.7
0.1	0.39	1.0	2.0	32	88.8

Fig. 8 — Effect of mechanical wearing-in of contacts.

6.6 *Effect of Non-Electrical Operation of Contacts*

The reduction in failure rates that is found on relays with a previous operational history suggests that the contacts "wear in" mechanically, thereby improving the contact fit and changing the surface roughness. To check this hypothesis, four groups of nine test relays each were operated at rates of one or ten operations per second without wires on the contacts. One of these groups was operated for one million operations and three for ten million operations. Each of these groups of relays were then wired into an OCTM, the contacts being replica-cleaned in each case. It was found that the failure rate in these four tests averaged about one-half the failure rate obtained for nine other tests using relays which were not subjected to an earlier wear-in. The data are shown in Fig. 8. This difference, however, may not be statistically significant because of the large variability in the data. The improvement in contact performance that is obtained with contacts that have had a previous operational history also seems to be associated with electrical operation of the contacts. This effect is explored in the succeeding sections.

6.7 *Effect of Charging of Wire Capacitance on Contact Closure*

Under some circuit conditions, contacts close without potential difference and open without current. Such contacts are electrically non-

eroding. Under other circuit conditions, the operation may differ only to the extent that the contacts, upon closure, charge the capacitance of a few feet of wire. Such contacts are also classed as non-eroding contacts, although it is obvious that minute erosion must take place, inasmuch as a potential difference exists when the contacts close. For brevity, contacts of the first type are called "no cable charge" (NCC) contacts and contacts of the second type are called "cable charge" (cc) contacts. From a failure standpoint, the NCC contacts are appreciably worse than the cc contacts.

To determine the effect of cable charging closure upon the failure rate, an OCTM was operated with the contacts of five AF relays connected without cable charging on closure and the contacts of four other relays connected with cable charging. The failure data are plotted in Fig. 9. For the first million operations, the NCC contacts had 3.3 times the failure rate of the cc contacts. At one million operations, the contacts were all replica-cleaned and the NCC contacts were changed to cc operation, so that all contacts were then connected alike. The failure rate dropped for both groups and became more nearly equal, but the contacts with the early history of NCC operation still showed a 50 per cent higher failure rate. At three million operations, the contacts were again replica-cleaned and reconnected as they had been initially. The failure rates again dropped, this time to insignificant failure rates for each group.

There seem to be two distinct explanations for the effect of the electrical circuit in reducing the failure rates noted in this test. On the one hand, the capacitance-charging current on closure appears to burn and carbonize the polymer, thereby reducing the open-contact failure rate.
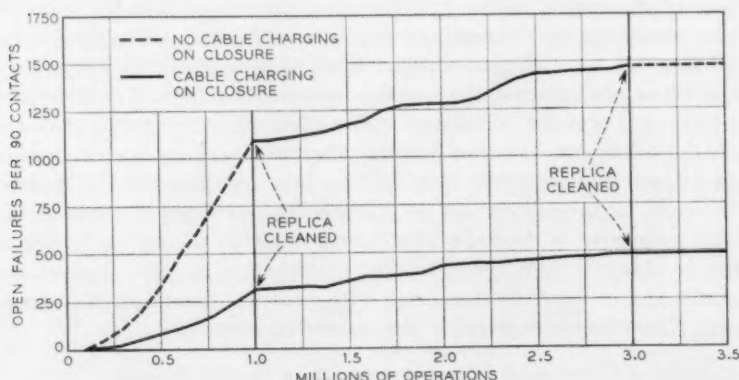

Fig. 9 — Effect of charging of wiring capacitance on contact closure.

However, the capacitance-charging current appears also to have a permanent effect on the contact surfaces, since the failure rates drop abruptly and remain permanently lower after each cleaning. Even in the case of contacts having no cable charging on closure, some arcing takes place when the polymeric films are punctured by the 50-volt checking voltage. This arcing alone is enough to improve the contact performance, being almost as effective as the cable-charging currents on closure.

It is apparent that this type of arcing produces a permanent improvement in the contact performance. The arcing burns off the larger irregularities on the contact surfaces, thereby tending to increase the fit or alignment of the contacts. It also increases the microscopic roughness, increasing the number of potential conducting areas and possibly reducing the failure rate from fine particles, by the mechanism described by J. B. P. Williamson, J. A. Greenwood and J. Harris.[10]

A further attempt to explore the effect of arcing on the contact performance is described in the next section.

6.8 Effect of Contact Erosion

Precious metal contacts, when used in a circuit arrangement that results in contact erosion, have a very low failure rate from dirt or organic films. For example, during earlier laboratory contact erosion studies in which thousands of contacts were operated millions of times, the failure rate from dirt and films was only 0.01 failure per 10^9 contact operations — a remarkably low failure rate. The obvious explanation is that heavy arcing burns away dirt and films thereby keeping the contacts clean.

The preceding section, however, indicates that the arcing may also produce a permanent improvement in the contact surfaces. To test this possibility, a group of nine AF relays with a total of 90 single contacts was operated in a sealed container, each contact opening and closing a non-inductive load of 2500 ohms, with a 50-volt battery. The relays were operated 80,000 times each, at which time it was evident by oscilloscope observations of the increased arcing that the contacts had become severely activated from organic vapors.[3] The erosion due to this treatment was insignificantly small because of the low number of operations. However, because of the presence of organic vapors, the arcing eroded the contacts over a large area with microscopic roughness, but without large buildups and pits. At the end of this priming procedure, half of the contacts were replica-cleaned and the 90 contacts were then operated in the OCTM without cable charging in the usual fashion. The failure rates for these slightly eroded contacts were found to be one-

tenth to one-fifth of the values normally found for new AF relay con-
tacts, the cleaned contacts being twice as good as the uncleaned con-
tacts.

This experiment was repeated using contacts that were roughened
in the open air by the erosion that occurs from 1000 operations of an
inductive relay load. This type of erosion produced larger pits and build-
ups on the contact surfaces and did not improve the fit. The failure rates
were found to be as high as for new untreated contacts.

These many experiments show consistently that polymer failures,
and probably failures from very fine dusts, will be minimized if the
contacts are large in area, well aligned so as to provide a good contact
fit and having surface roughnesses limited to microscopic size.

6.9 *Effect of Relay Coil Temperature*

When AF- or U-type relays with palladium contacts are operated
without enclosures on duty cycles so low as to produce essentially no
temperature rise in the coils, the failure rates are acceptably low. How-
ever, if appreciable coil heating takes place, the vapors given off by
the organic insulating materials in the relay increase the contact failure
rate enormously.

Fig. 10 shows the effect of relay coil temperature on the failure rate
of AF-type relays with twin contacts. These data were obtained in
six separate tests, in which more than 1100 contacts are represented.
Most of the tests ran a year or more; one test ran for 29 months. The
graph indicates that the failure rate at 68°F is one open per 10^9 contact
operations, and that the failure rate doubles for every 10°F increase
in coil temperature. A relay coil at 135°F indicates a rate of 100 failures
per 10^9 contact operations. Such coil temperatures are easily achieved
in telephone switching common control circuits: a 700-ohm relay with
a 25 per cent duty cycle or a 270-ohm relay with a 20 per cent duty cycle,
each in an ambient of 100°F, would produce such coil temperatures
and corresponding trouble rates.

In these tests, the air in the vicinity of the contacts became warmer
and drier at the higher coil temperatures, and this is possibly an ad-
ditional factor affecting the failure rates. The relative importance of
these factors has not been resolved.

6.10 *Effect of Relay Insulating Materials*

In the earlier investigation of the effect of organic vapors on contact
erosion,[3] it was found that all the relay insulating materials, including
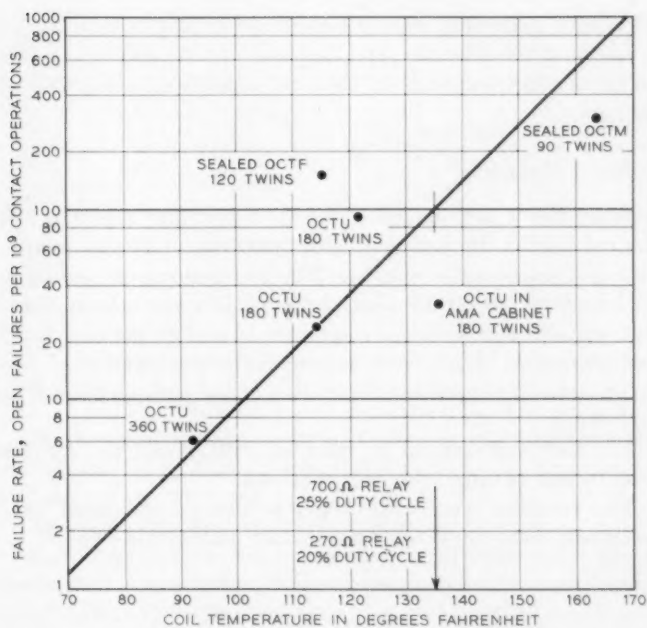
Fig. 10 — Effect of relay coil temperature.

the oleo-resinous wire insulation, the cellulose acetate sheets used in the coil construction and the phenol fiber spoolheads and insulators, were important sources of vapor. It was necessary to replace all these materials by inert materials (glass insulated wire, glass spoolheads, glass insulators, etc.) to prevent vapor contamination of the contacts.

The insulating materials used in modern relays are desirable for economy, freedom from corrosion and good dimensional stability with varying temperature and humidity. Other insulating materials may be found which have all these virtues and which also provide freedom from organic contamination of relay contacts, but such materials are not yet available.

On the AF-type relay, the plastic contact covers and the phenol fiber armature cards enclose the relay contacts. This design differs from the U relay design, which does not use contact covers. To determine the effect of the proximity of these parts on the contact failure rate, an OCTU with AF-type relays was operated using metal armature cards and metal contact covers. No significant reduction in failure rate

was obtained, indicating that the card and contact cover are not the major causes of the difficulty, and that no improvement can be expected by the use of other materials for the card and contact cover that enclose the contacts.

6.11 *Effect of Humidity*

Humidity has a pronounced effect on the failure rate and the persistency of failure. In one test, nine 2500-ohm AF-type relays were tested in the sealed OCTM chamber. The test was run at normal uncontrolled humidity (40–60 per cent) for 1,420,000 operations. Thereafter the test was alternated abruptly between 15 and 85 per cent humidity, using a test period of one week or more for each humidity.

The procedure followed for the cycling of the high and low humidity was as follows:

(1) The test was stopped at noon on Friday and the conditioning material (water or silica gel) was removed.

(2) The chamber was dried, when necessary, and aired out. The relay contacts were protected against dust during this step.
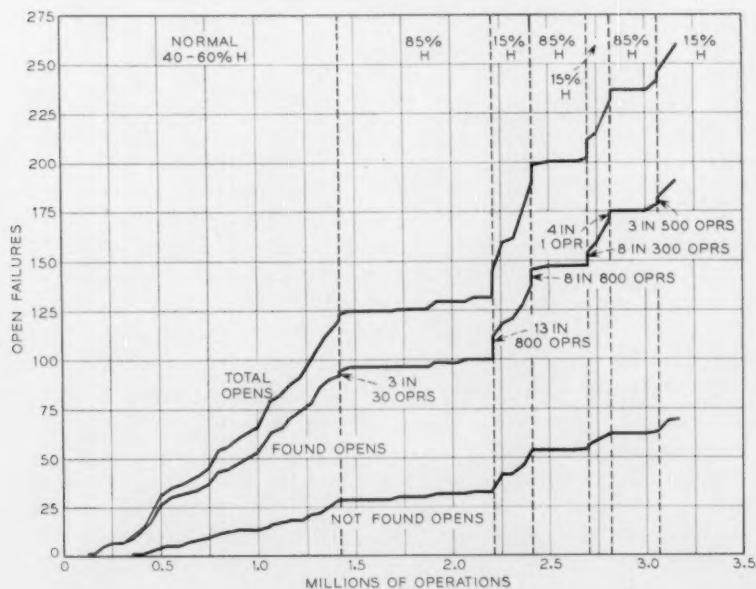

Fig. 11 — A plot of open failures against operations, showing the effect of humidity.
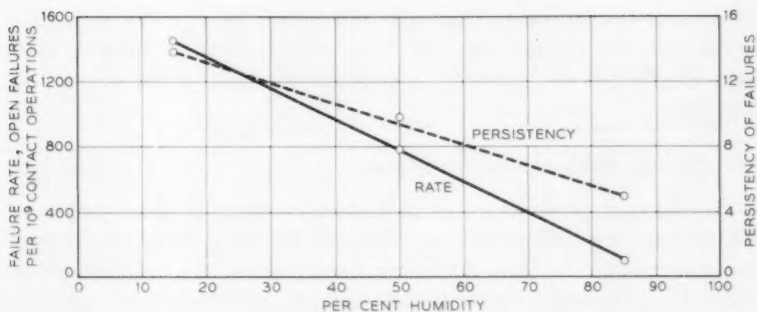
Fig. 12 — A plot of open failures and the persistency of those failures against humidity.

(3) The new conditioning material (water or silica gel) was placed in the chamber. The chamber was sealed and the atmosphere allowed to stabilize over the week-end.

(4) The week's test was restarted on Monday morning.

A plot of open failures against number of operations for the entire test is shown in Fig. 11. Three curves are shown: (1) found opens, (2) not found opens and (3) total opens. Also shown are the humidity values for each part of the test.

When the test was restarted for each new condition of humidity, a flurry of failures occurred within the first 1000 operations or less. These failures are also shown in Fig. 11. The failures occurring at these transition periods are possibly due to a shift in the actual point of contact resulting from a slight physical change in the relay's insulating parts caused by the change in humidity. The failures may also be partly due to a change in the volume of the polymer which had caused it to enter the contact area.

Discounting the flurries of opens that occurred during the transition periods, the failure rates for "wet", "normal" and "dry" humidity conditions are 100, 780 and 1340 opens per billion contact operations. Corresponding persistencies of these failures are 4.9, 9.8 and 12.8 respectively. A plot of these rates and persistencies is shown in Fig. 12. It is interesting to note that there is a linear relationship between humidity and rate of failure, and between humidity and persistency: the lower the humidity the higher the rate of failure and persistency. The following explanation appears reasonable. The polymer takes on more water as the humidity increases, becoming less dusty and less likely to fall into the contact area. Conversely, the drier the polymer

becomes at low humidity, the better the chance for powdery aggregates to fall into the contact area and thus produce contact failures. Tests have shown that the quantity of polymer generated is not affected by humidity.

6.12 *Effect of Relay Design Variations*

The contact performance of palladium contacts in the presence of organic vapors is intimately connected with the relay design and contact spring actuation. In the OCTF a direct comparison was obtained of the M24-, U-, UB-, and AF-type relays (Fig. 13), each with single and twin contacts (Fig. 14). This test was operated for 29 months. The contacts, in all cases, closed without cable charging. The failure rates listed in the Table VI remained essentially the same after the first few months. The actual data are plotted in Figs. 15 and 16.

The M24, U, and UB relays each have 26 grams contact force per twin contact, whereas the AF relay has 13 grams. The M24, U, UB, and AF relays have, in the order named, decreasing contact slide during contact closure. It is believed that the remarkable difference in contact failure rates for these relays is due to these two factors. The peculiar
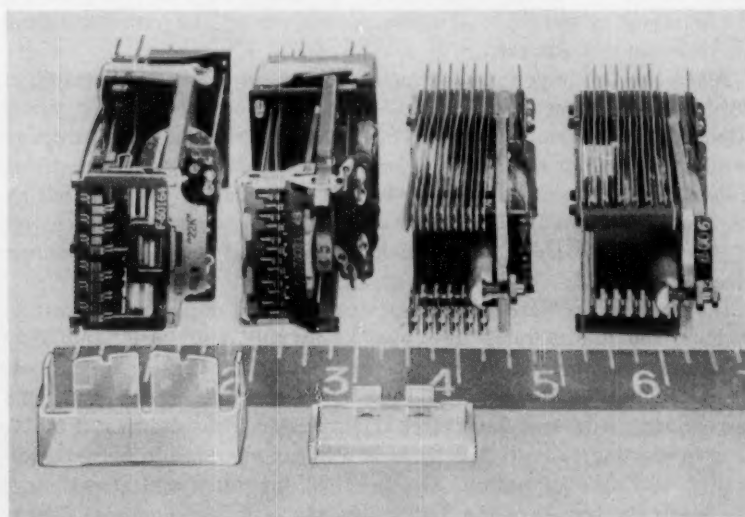

Fig. 13 — From left to right: wire-spring relays AF and M24, U- and UB-type relays. In foreground are individual plastic contact covers for the AF and M24 relays, respectively.

TABLE VI — FAILURE RATES WITH DIFFERENT RELAY TYPES

Relay	Single Palladium Contacts		Twin Palladium Contacts	
	Opens per 10^9 Contact Operations	Persistency	Opens per 10^9 Contact Operations	Persistency
M24	24	4.0	1	20.5
U	400	7.6	5	11.9
UB	2570	10.8	32	18.3
AF	4150	6.3	148	14.8

relations between the single- and twin-contact failure rates and persistencies are discussed in another section.

6.13 *Effect of Contact Velocity*

The magnitude of the contact-closing velocity has a significant effect on the failure rates of palladium contacts. As the velocity is increased, the impact and resulting slide is increased with a corresponding decrease in the open-contact failure rates.

It is found in the octu and octf tests of the AF relays that normally-open single contacts (makes) fail about three times as often as normally-closed single contacts (breaks). Similarly, normally-open twin contacts fail about ten times as often as normally-closed twin contacts. Since the contact actuation of normally-closed and normally-open contacts is the same on the AF relay, and since these relays are operated on about a 50 per cent duty cycle providing equal closed and open intervals, it
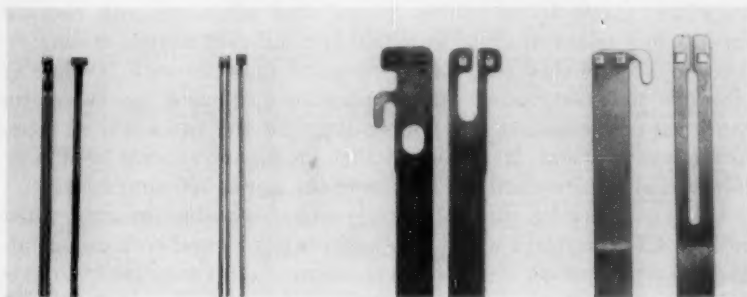

Fig. 14 — Mating twin contact pairs of AF, M24, U and UB relays. The AF and M24 twin contacts are coined to a cylindrical shape, the other contacts have flat surfaces. The AF and M24 twins are made up of two singles which are relatively closer together than in the U and UB relays. Also, the AF and M24 have a common fixed contact, while the U and UB have individual contacts.
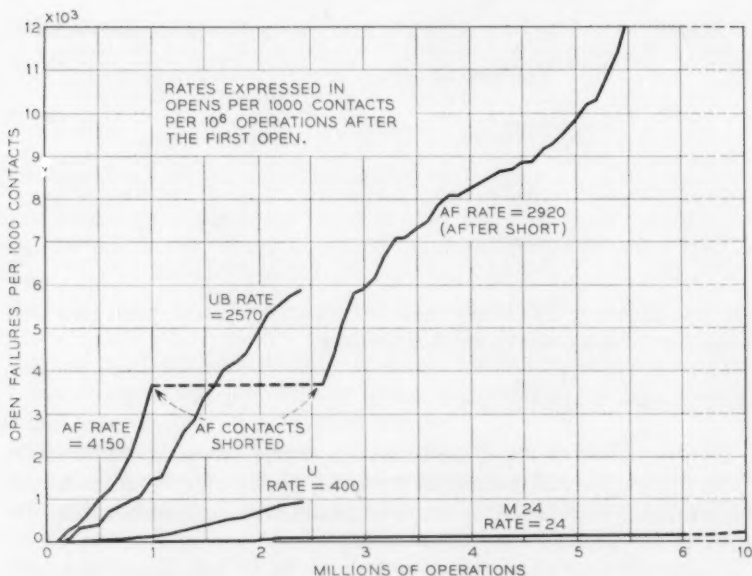
Fig. 15 — A plot of single-contact open failures against operations for AF, M24, U and UB relays operating in the octf test.

is concluded that this difference in failure rates must be due to a difference in contact-closing velocity.

A check on this hypothesis has been obtained in the octm by comparing contact failure rates on relays which have different contact-closing velocities. Contacts on 16-ohm relays were compared with contacts on 700-ohm relays in order to obtain two different contact-closing velocities. The 16-ohm relays were operated in series with 90-ohm resistances mounted outside the compartment, thereby providing the same power dissipation and temperatures for the two kinds of relays in the compartment. It was found that the higher velocity 16-ohm relays had about one-third the failures of the slower 700-ohm relays.

Other evidence for effect of velocity was obtained in the octf, where certain AF-type relays which were deliberately slowed both on operate and release by means of series and shunt resistances were found to have the very high failure rate of 2610 opens per 10^9 contact operations. This failure rate was the highest steady-state failure rate experienced on twin contacts in the entire test program. When the series and shunt resistances were removed to obtain high velocity, the failure rate jumped to the enormous figure of 69,000 opens per 10^9 operations during the next
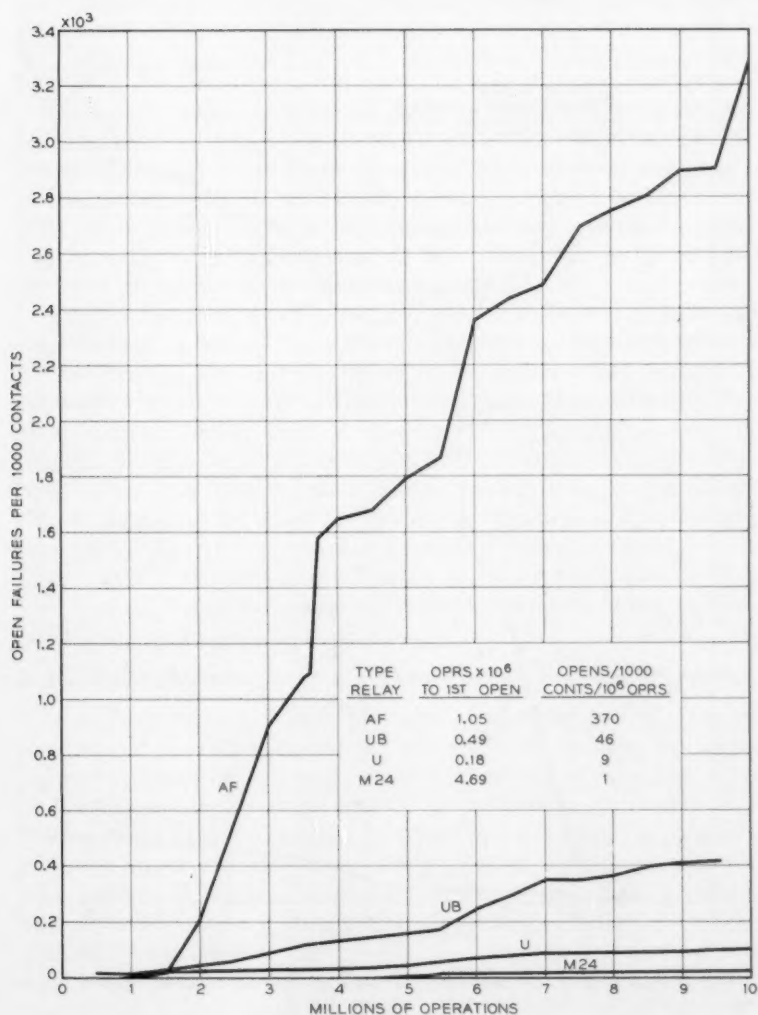
Fig. 16 — A plot of twin-contact open failures against operations for AF, M24, U and UB relays operating in the octf test. The failure rates are higher than those shown in Table VI because the low-velocity relay data described in Section 6.13 are included.
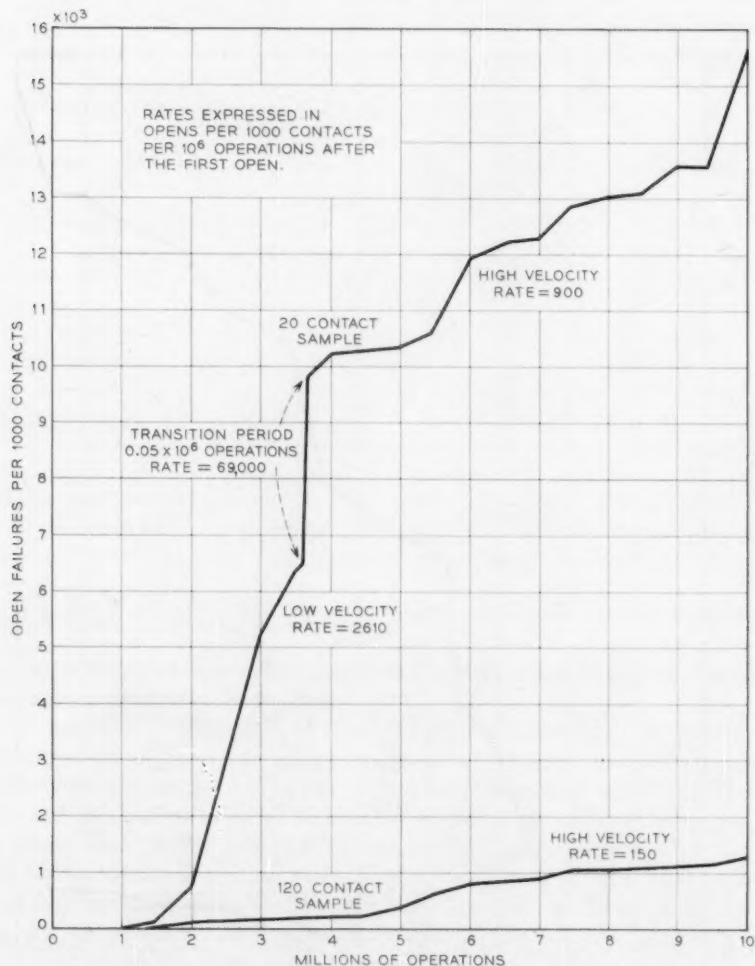
Fig. 17 — A plot of AF wire-spring relay twin-contact failures against operations, showing the effect of contact-closing velocity.

50,000 operations before settling down to the lower rate of 900 opens per 10^9 operations. It appears that the lower velocity and therefore lower shock of the low-velocity test relays allowed a heavy accumulation of polymer in the vicinity of the contact and thereby increased the failure rate. When the velocity was abruptly increased, the polymer was disturbed, fell into the contact area and temporarily increased the failure

rate. Comparison relays with high velocity in the same test had a failure rate of only 150 opens per 10^9 operations. The data are shown in Fig. 17.

6.14 *Effect of Vibration*

External vibration is a serious cause of contact failures on relays having normally closed palladium contacts. The vibration, provided it is severe enough to overcome the static friction of the closed contacts, causes a slight contact slide which, in the presence of organic vapors, produces polymer very near the true contacting area. The contacts may then fail in the same fashion as operating contacts, by producing an open, low resistance or noisy transmission.

The vibration effect can be surprisingly large. In several tests using single palladium normally-closed contacts on wire-spring relays subjected to vibration and vapors from a large number of other wire-spring relays, more than one-third of the contacts were found to register open circuits after several hundred hours of the test. When the contact force was doubled, or when the closed contacts carried a continuous current to burn and carbonize the polymer, the failure rates were found to drop sharply. The contacts had the usual low-persistency characteristic of operating contacts in that a few mechanical operations cleared the open.

As was the experience in the relay operating tests, 22-karat gold, No. 1 metal, No. 3 metal and silver performed much better than palladium in the presence of organic vapors.

6.15 *Relation Between Single and Twin Failure Rates and Persistencies*

If each contact of a twin pair were electrically and mechanically independent of each other, then the failure rate of the twin pair, neglecting its persistency of failure on successive operations, would be related to the failure rate of a single contact by the formula $F_T = (F_s \bar{P}_s)F_s = F_s^2 \bar{P}_s$, where F_s is the single failure rate, not including its persistency of failure on succeeding operations and $\bar{P}_s$ is the average persistency of such failures. Similarly, the average persistency of failure of the twin pair would be related to the average persistency of failure of the single contact by the formula $\bar{P}_T = \frac{1}{2} \bar{P}_s$.

These formulae completely fail to predict twin contact performance from the data taken on single contacts. In the OCTF the AF relay, with halves of the twin contacts removed to provide single contacts, had a failure rate of 4150 opens per 10^9 operations and an average failure persistency of 6.3. The above formulae would predict a twin failure rate of 0.1 failure per 10^9 operations and an average persistency of 3.15. The twin contacts in the same test had an actual failure rate of 148 opens per 10^9 operations and an average failure persistency of 14.8, showing that the formulae do not apply.

The failure of the simple formulae to apply can be ascribed to the fact that twin contacts are neither electrically nor mechanically independent. From an electrical standpoint, one half of the twin pair will inevitably be more vulnerable to carbonization of the polymer film by the testing voltage and current. The other twin contact, shunted by its lower-resistance twin, will not suffer electrical degradation of the polymer, since it will never be subjected to a voltage breakdown and will never carry the electrical current unless the other twin has failed. Consequently, the failure probabilities of the two halves differ greatly and the twin failure rate cannot be easily determined from data on single contacts. Also, from a mechanical standpoint, the twin contacts are not independent, since polymer generated on either twin may contaminate the other. The amount of such mutual contamination depends on the proximity of the twins and whether the fixed contact is a single-bar contact, as in the AF- and M24-type relays, or two twin bars as in the U and UB relays.

A number of tests have indicated that this lack of electrical and mechanical independence is the cause of the high twin failure rates. Additional probing tests are being made to investigate this effect more completely.

Empirical formulae which relate the single and twin failure rate actually measured in the OCTF and the OCTM are as follows.

For wire-spring relays AF and M24, $F_T = 1/28\ F_s$, and for U and UB relays, $F_T = 1/80\ F_s$. The AF and M24 twins are made up of two singles relatively closer together than in the U and UB relays. Also, the AF and M24 have common fixed contacts, whereas the U and UB relays have individual fixed contacts. These differences are apparent in Figs. 13 and 14 and appear to explain the higher failure rates of twins relative to singles on the AF and M24 relays.

6.16 Low-Resistance Characteristics

In the 50-volt, 20-milliampere, automatic checking circuits used in these tests, the contact resistance is always found to be less than 20 ohms or more than 100,000 ohms. If the insulating film has sufficiently high intrinsic resistance and sufficient dielectric strength, practically no conduction takes place. Under such conditions, very high resistances above 100,000 ohms are measured, and these are referred to as "open" contacts. However, if dielectric breakdown by ionization from the voltage or thermal breakdown from the current occur, the power dissipation in the contact surfaces will be too great at intermediate contact resistances, and breakdown will proceed until the lower range of contact resistance results. Of course, if the circuit voltage is kept low or if the circuit impedance is kept high, the power dissipation at the contacts

will be limited and intermediate values of resistance would then be expected.

The low-resistance effect, unlike the random open contact failures, tends to remain fairly consistent from operation to operation throughout the test. In general, a particular contact that develops the low-resistance effect does not produce open contacts, and vice versa. For example, palladium contact tests on U-type relays in the OCTM consistently showed that approximately one-third of the contacts would develop low resistance, one-third would fail randomly from opens and one-third of the contacts would produce neither low resistance nor open contacts. Gold and gold-silver alloys show only the low-resistance effect, but not on all contacts in a given test; the others have practically no resistance. The distribution of resistance measurements made on various contact metals on U-type relays in OCTM units during open-contact test runs is shown in Table VII. All resistance measurements were made with 1.5 volts and 40 to 80 milliamperes. High resistance measurements that occur with open contacts are not included.

TABLE VII — DISTRIBUTION OF CONTACT RESISTANCE WITH VARIOUS METALS. SINGLE CONTACTS ON U RELAYS

Contact Metal	Relay Operations in Millions	Number of Contacts			Maximum Resistance in Ohms
		In Test	>0.1ω	>1ω	
Pd/Pd	2.5	90	34	29	8.5
No. 1/No. 1	10.0	90	18	10	5.0
No. 3/No. 3	10.5	90	20	13	5.5
22K Au/22K Au	10.0	60	33	27	17.5

These contacts were essentially non-working electrically. However, they did charge the wire to the succeeding contact in the chain to 50-volt potential as they closed and, to this extent, they were electrically working contacts. To determine the effect of this small amount of electrical erosion in carbonizing the polymer, 30 additional contacts in the 22-karat gold test listed above were tested simultaneously without such cable charging, and it was found that only three of these contacts had resistance greater than 0.1 ohm and only one contact exceeded 1 ohm.

The wire-spring relay test results were similar except that a higher percentage failed by high resistance (opens) and a correspondingly lower percentage showed low resistance. The effect of cable-charging closure was to increase the percentage that showed the low-resistance effect, as for the U relays. Also, higher-velocity wire-spring relays had a much greater percentage of contacts with low resistance, again indicating that high impact forces and high slide carbonize the polymer.

6.17 *Noise on Speech Transmission Contacts*

Undesirably high noise levels were encountered in the speech transmission circuits of the No. 5 crossbar system due to deposits of polymer and carbon on the normally-closed palladium contacts of certain relays on the trunk link frame. The noise occurred only when nearby relays released, jarring the frame and causing the contaminated contact to become microphonic. The noise at times sounded to the listener like a banjo twang of such a magnitude that words or even sentences occasionally became unintelligible.

Field tests on U-type relay equipments in service showed that 30 per cent of the noise-producing relays in a large sample would produce noise exceeding 25 db* (just audible). When the contact metal was changed to No. 1 alloy, only 1 per cent of the contacts had objectionable noise. When soft rubber vibration mounts were used with the original contact metal, the number of contacts with audible noise was 5 per cent.

6.18 *Mechanical Wear*

Life tests to determine the mechanical life of 22-karat gold overlay contacts on wire-spring relays indicate a life exceeding 100,000,000 operations. These tests, run with unwired contacts at an accelerated rate, show considerable variation between contacts and between tests with respect to the amount of wear. It is believed that the contacts obtain some lubrication from the products formed from the organic vapors and that, with the slower operating rates that occur in practice, the mechanical life of the gold overlay will exceed the life indicated in these accelerated tests.

Although relays in circuits such as markers and Accounting Center machines may operate a billion times or more during their lifetimes, no trouble is expected with these relays after the gold wears off because of the often-repeated laboratory indication that relays with high operating rates and a history of many operations have low failure rates.

6.19. *Final Tests*

To test the final product, 90 standard AF relays with the gold overlay on the moving contacts were tested in an OCTU, which most closely resembles service conditions. All contacts had been cleaned prior to start of the test, since the present manufacturing process calls for thorough cleaning of the contacts by wet-scrubbing with trichlorethylene.[11] No contact failures occurred in a year's operation of this test. Identical tests made on palladium contacts without gold overlay have had about 60 found open contact failures in this time.

---

* 25 db above 10^{-12} watts

VII. CONCLUSION

The obvious solution to many of the difficulties experienced with precious metal contacts is to eliminate organic vapors from the contact environment. However, this is often not feasible. The plastics and phenolics used in relay insulation are potent sources of these undesirable vapors, but these insulating materials are desirable for other reasons. Forced ventilation with adequate filtering to remove dust would be uneconomical, as would be the use of activated charcoal to absorb the vapors. For the telephone plant, where so many relays and contacts are used, the best solution would be to make the contacts immune, or at least less susceptible, to the vapors.

Palladium and other platinum-family metals are particularly susceptible to the vapor effects. These metals are suspected of acting as catalysts to form highly insulating films, and no poison has been found which will markedly reduce this effect. The best solution found so far is to overlay the palladium with a gold alloy sufficiently thick to provide for mechanical wear. Where a solid, homogeneous contact without gold overlay is preferred and where the added cost can be justified, Western Electric No. 1 metal can be used and is practically as good. However, although the high resistance effect is eliminated, these gold or gold alloy contacts can still be expected to show the low-resistance phenomenon in the presence of organic vapors. Also, the use of gold does not reduce the heavy erosion that takes place on electrically working contacts in the presence of organic vapors.

Gold overlay on only the moving contacts of wire-spring relays is used for economy and ease of manufacture. The good performance obtained is the result of the decreased polymer formation and the twin contact action coupled with the low persistence of failure characteristic of the polymer. Where the polymer difficulty has occurred on single-contact relays, as in step-by-step systems, no compromise is made and the change to gold alloy (No. 1 metal) is made on both mating contacts.

Preliminary reports from early installations indicate that the contact performance of the new wire-spring relays with gold overlay on the moving contacts will be highly satisfactory.

VIII. ACKNOWLEDGMENTS

on contacts was done by H. W. Hermance and T. F. Egan, to whom we are indebted for many stimulating discussions. The effect of contact vibration was discovered and explored by the late C. E. Nelson. The electrical breakdown characteristics of the polymer were investigated by Miss P. M. Hamway. Introduction of the practical results of these tests into relay manufacture was coordinated by H. N. Wagar.

APPENDIX

Simplified Schematic of the OCTM (Fig. 18)

When the start key is operated, and later released, the AL relay operates, which in turn operates the ST relay (both these relays remain operated until a contact failure occurs). The operation of the ST relay passes interrupted ground pulses into pulse-dividing relays W and Z. The first ground pulse operates the W relay, which in turn operates
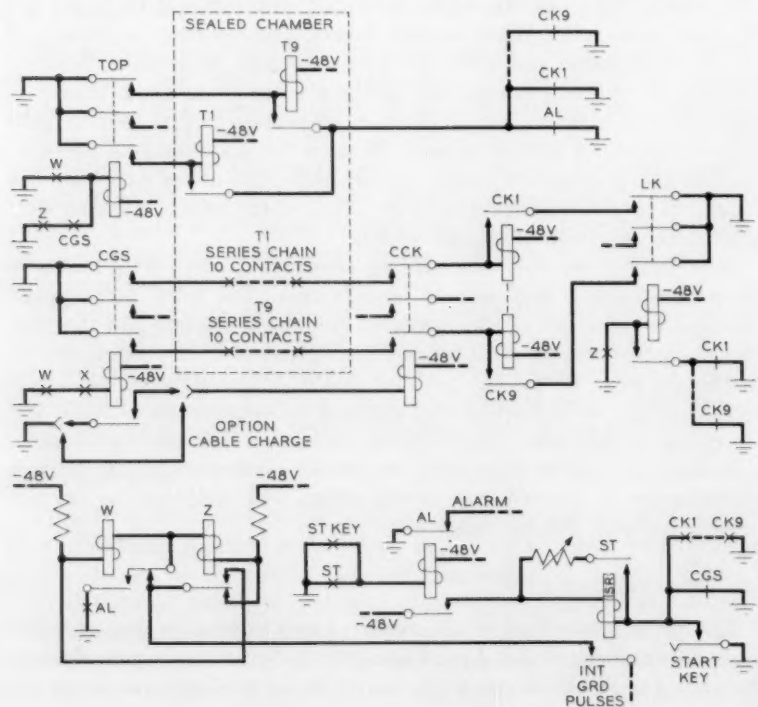
Fig. 18 — Simplified schematic of the ocтм.

the TOP relay, and hence the nine test relays Tl through T9, preparing the contact chains for checking. The first open pulse operates the Z relay, which operates the LK and CGS relays. The CGS relay operates the CCK relay and supplies a ground to the contact chains. If all test contacts are low in resistance, the corresponding check relays CK1 through CK9 operate and lock up, returning the test contacts to ground potential. The second ground pulse releases the W relay, which in turn releases the CGS, which then releases the CCK and TOP relays. The TOP relay releases the T relays. The second open pulse releases the Z relay, which releases the LK relay. The LK relay in turn releases the CK relays, thereby completing the cycle. When a failure occurs on a test contact, a check relay CK will fail to operate. The failure of a check relay allows the ST relay to release, stopping and holding the circuit in the failed position and releasing the alarm relay AL.

Fig. 19 — Simplified schematic of the OCTU.

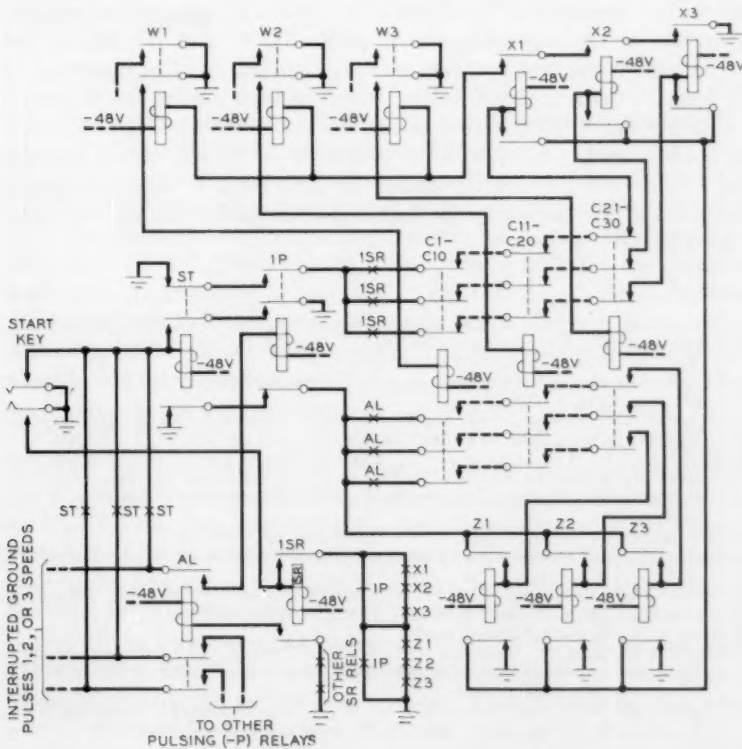*Simplified Schematic of the OCTU (Fig. 19)*

Approximately one-third of the circuit is shown in the schematic diagram. The operation of the start key operates the ST and 1SR relays. The 1SR relay operates the AL relay, which allows interrupted ground pulses to reach the pulsing relay 1P (the AL and 1SR relays remaining operated until a contact failure occurs). The release of the start key allows the ST relay to release when all three interrupted ground pulses simultaneously reach a "no ground" period. (The ST relay remains unoperated during the test.) The first operation of the pulse relay 1P, after the release of the ST relay, supplies a ground to the break-contact series chains of test relays C1 through C30. If all test contacts are low in resistance, the corresponding check relays X1, X2 and X3 operate and lock up, returning the test contacts to ground potential. The operation of all X relays operates the W relays, which in turn operate the C relays. The operation of the C relays prepares the make-contact chains, operating the corresponding check relays Z1, Z2 and Z3, which lock up, returning the test contacts to ground potential. The operation of all Z relays releases the X relays, which in turn release the W relays, and hence the C relays, completing the cycle. When a failure occurs on a test contact, a check relay (either X or Z) will fail to operate. The failure of a check relay allows the 1SR relay to release, stopping and holding the circuit in the failed position and releasing the alarm relay AL.

REFERENCES

1. Watson, E. A., Proc. of Inst. of Automobile Engineers, **22,** 1927–28, p. 812.
2. Chaston, J. C., Materials for Electrical Contacts, J. Inst. Elec. Engrs. (London) **88,** Part II, August 1941, p. 276.
3. Gumley, R. H., Contact Phenomena in Sealed Containers, Bell Labs. Record, **32,** June 1954, p. 226.
4. Germer, L. H. and Smith, J. L., Activation of Electrical Contacts by Organic Vapors, B.S.T.J., **36,** May 1957, p. 769.
5. Hermance, H. W. and Egan, T. F., Examination of Contacts and other Surfaces by a Plastic Replica Technique, in *Proceedings of 1956 Electronic Components Symposium, Washington, D. C.*, Engineering Publishers, N. Y., 1956, p. 85.
6. Hermance, H. W. and Egan, T. F., this issue, pp. 739–776.
7. Keefer, H. J., Dust on Relay Contacts, Bell Labs. Record, **35,** January 1957, p. 25.
8. Holm, Ragnar, *Electric Contacts*, Almquist & Wiksells, Stockholm, 1946. *Electron Contacts Handbook*, Springer Verlag, Berlin, 1958.
9. Knapp, H. M., unpublished work.
10. Williamson, J. B. P., Greenwood, J. A. and Harris, J., The Influence of Dust Particles on the Contact of Solids, Proc. Roy. Soc. (London), Series A, **237,** November 20, 1956, p. 560.
11. MacDonald, R. W. and Hermance, H. W., A New Method for Cleaning Wire Spring Relays, Bell Labs. Record, **36,** January 1958, p. 30.

Gray Codes and Paths on the n-Cube

By E. N. GILBERT

Certain problems in coding and in switching theory require a list of distinct binary n-tuples such that each differs from the one preceding it in just one coordinate. Geometrically, such a list corresponds to a path which follows edges of an n-dimensional cube. This paper finds all types of closed paths on cubes with $n \leqq 4$. For larger n, a process given here will produce large numbers of paths.

I. INTRODUCTION

A Gray code is a means of quantizing an angle and representing it in a binary alphabet. The encoding is such that angles in adjacent quantum intervals are encoded into n-tuples of binary digits which differ in just one place. For example, taking $n = 3$, as the angle increases from $0°$ to $360°$, the binary code for the angle might go through the succession 000, 001, 011, 010, 110, 111, 101, 100 and back to 000. Gray codes are used when the encoding is performed by a code wheel. At angles close to the boundary between two quantum intervals, any of the digits which change at the boundary are likely to be in error. In a Gray code there is only one such questionable digit, and a mistake in this digit only gives to the angle the code for the adjacent quantum interval.

Although the Gray code example given for $n = 3$ is easily generalized to obtain the well known Gray (reflected binary) code for any n, there are, in general, a large number of other codes which also change one digit at a time. Our problem is to find these other codes. In special applications, some of the others may be preferable to the conventional Gray code. For instance, it may be desirable to use other numbers of quantum intervals beside powers of 2; powers of 10 might be a natural choice. If the quantity being encoded is a length rather than an angle, one can drop the requirement that the first and last quantum interval have codes differing in just one position. Then there is a still larger variety of encodings from which to choose.

For $n \leqq 4$, we will exhibit all possible codes which recycle (i.e., are

815

suitable for angles). As n grows beyond 4, the number of possibilities soon becomes enormous; we will give a procedure by which a large number of codes may be constructed. The purpose of this paper is to construct, classify and catalogue codes, but no attempt will be made to single out those codes which are particularly useful.

In searching for codes the following geometrical picture is helpful. The set of edges of a unit n-dimensional cube forms a linear graph which we will call the *n-cube graph* Q_n. Q_n has 2^n vertices, each labeled by an n-tuple $(x_1, \cdots, x_n)$, $x_i = 0$ or 1. Two vertices are joined by a line of Q_n if their coordinates agree in all but one of the n places. We are interested in the paths and cycles of Q_n. By a *path of length* L is meant a set of lines $(V_0 V_1, V_1 V_2, \cdots, V_{L-1} V_L)$ of Q_n where the vertices V_0, $\cdots$, V_L are distinct. *Cycle of length* L is similarly defined but with $V_1, \cdots, V_L$ distinct and $V_0 = V_L$. If we follow a path or cycle of Q_n and interpret the n-tuples which label the vertices as codes for successive quantum intervals, we obtain an encoding in which only one digit changes from interval to interval. Thus, our problem is to find all paths and cycles of Q_n. The cycles are the encodings suitable for angles.

A permutation of digits in the n-tuples of a code produces a new code which is not significantly different from the original one. Its code wheel is obtainable from the original wheel merely by permuting tracks. Similarly, a complementation (interchange of 0 and 1 in certain coordinate positions of the n-tuples) is a minor change. These operations correspond to rotation and reflection symmetries of the cube. The symmetry group of Q_n is the hyperoctahedral group O_n of order $2^n n!$. The typical symmetry operation of O_n consists of one of the $n!$ possible permutations of the coordinates of the n-tuple, followed by one of the 2^n possible complementations. Two paths or cycles will be called *equivalent* or of the same *type* if one can be changed into the other by applying to Q_n one of the symmetry operations in O_n. Although there are tremendous numbers of paths and cycles, it suffices to give just one of each type. Even our list of distinct types of cycles becomes rather long at $n = 4$. The exact number of types of cycles is not known for larger values of n. We give below a procedure for constructing large numbers of paths and cycles. If $L = 2^n - 1$, we can specify in advance the numbers $N_1, \cdots, N_n$, where N_k is the number of times x_k changes as one follows the path from V_0 to V_L. A similar specification is possible for cycles of length 2^n. The cycles of this length contain every vertex and thus are the Hamilton lines of Q_n (Ref. 2, Ch. 2). Even the number of types of Hamilton lines is found to grow rapidly with n; there are nine types of Hamilton lines for $n = 4$.

Another application of paths on Q_n may be found in switching theory. Here the labels on the vertices represent the states of a relay network with n relays. Coordinate x_i is 0 or 1 according to whether the ith relay magnet is turned off or on. Since it is physically impossible to change the state of two relay magnets precisely simultaneously, the state of the entire network can change only by following lines of Q_n. If the network is a counting circuit it is made so that its state follows a path or a cycle according to whether the network is intended to lock or recycle at the end of the count (Ref. 1, Ch. 11).

II. COORDINATE SEQUENCES

A path is specified completely by listing the $L + 1$ vertices V_0, $\cdots$, V_L in order. For example, on the 3-cube, a list might be

$$000, 001, 011, 010, 110, 100, 101, 111.$$

Ignoring the starting vertex, a more compact notation is to list in order only the coordinate places in which the change occurs. In the example cited, one would obtain (3231232). This L-tuple of coordinate places will be called the *coordinate sequence* for the path.

The vertices of a path might equally well have been written down in the reverse order V_L, $\cdots$, V_0. Hence the list of coordinate places written in reverse order [i.e., (2321323)] will be regarded as another notation for the same coordinate sequence.

Each coordinate sequence represents not only the given path but also every other path obtainable from it by one of the 2^n complementations. All of these paths are of the same type; it suffices here to study coordinate sequences rather than the paths themselves.

Two paths, P and P', are of the same type if and only if one of the $n!$ permutations of coordinates changes the coordinate sequence of P into the coordinate sequence of P'. Our problem thus becomes one of classifying coordinate sequences into symmetry types with respect to the symmetric group on the n coordinates.

In a similar way, a cycle (say 001, 011, 111, 101, 001) may be represented by a coordinate sequence (2121). Now, however, in addition to the two orders in which the coordinates may be written, there are L different vertices at which the list of changing coordinates can begin. Thus, for any cycle there may be as many as $2L$ distinct L-tuples of digits, all of which are considered to be the same coordinate sequence.

Not every list of digits is the coordinate sequence of a path or a cycle. A typical list, such as (121323131), represents a way of wandering along

lines of the cube graph, possibly visiting some vertices more than once. In the example cited, the vertices visited after the first and the sixth steps are the same. In the intermediate steps (213231) all digits $k = 1$, 2, 3 appear an even number of times; for all k the net change in x_k during these steps is even, i.e., zero. This observation leads simply to the following result.

Theorem I. An L-tuple $A = (a_1, a_2, \cdots, a_L)$ $a_i = 1, \cdots, n$ is the coordinate sequence of a path of Q_n if and only if every one of the $L(L + 1)/2$ blocks of consecutive digits $(a_i, a_{i+1}, \cdots, a_j)$ contains at least one of the n digits an odd number of times. A is the coordinate sequence of a cycle of Q_n if and only if every one of the blocks of length $1, \cdots,$ or $L - 1$ contains some digit an odd number of times while A itself contains every digit an even number of times.

As an illustration of Theorem I, we construct a simple cycle of length 2^n (Hamilton line) on Q_n. For $k = 1, 2, \cdots, 2^n$ define the kth digit of the 2^n-tuple to be the number

$$a_k = \operatorname*{Max}_{1 \le d \le n} (d \text{ such that } 2^{d-1} \text{ divides } k).$$

That one obtains a cycle thereby is easily proved by induction on n. When $n = 2$, the construction yields (1212), the coordinate sequence of a cycle on the square. For larger n, the construction yields a 2^n-tuple of the form

$$A_n = (B_n, n, B_n, n), \tag{1}$$

where B_n is the $(2^{n-1} - 1)$-tuple obtained from A_{n-1} by removing the $n - 1$ in the 2^{n-1} place. Consider any block C of $q < 2^n$ consecutive digits of A_n. If C does not contain one of the two n's, it is a block of the path B_n. If C contains only one n, n appears an odd number of times. If C contains both n's, the digits of A_n not in C form a block of B_n and contain some digit k an odd number of times; since k appears an even number of times in A_n it appears an odd number of times in C. In any case, some digit appears an odd number of times in C. The Hamilton line in question corresponds to the conventional Gray code.

III. THE 4-CUBE

Since cycles are of greatest interest, we have constructed a list of all types of cycles on Q_4. This list also includes all the cycles on Q_2 and Q_3, since such cubes are contained as subgraphs (faces) of Q_4. There are no cycles of odd length, since every coordinate must change an even number of times. The numbers of types of cycles of lengths $4, 6, \cdots, 16$

are 1, 2, 7, 10, 23, 20 and 9. Of these, the cycle of length 4 is found on Q_2; both cycles of length 6 and one cycle of length 8 are found on Q_3.

The computation consisted of the enumeration of a large number of cases using whatever *ad hoc* simplifications could be found. For example, consider any cycle of length L which contains no pair of points which differ in all four coordinates. Complementing all four coordinate places changes this cycle into a new disjoint one. Then $2L \leq 16$. It follows that every cycle of length 10 contains such a pair of "diametrically opposite" points. The cycle can be cut into two paths joining these points, one of length 4, the other of length 6. Hence, coordinate sequences for types of cycles of length 10 can all be written in the form (1234 . . . . . .), and only the ways of filling the six empty places must be enumerated.

It is also helpful to draw a picture of the 4-cube in such a way that certain equivalences between cycles becomes geometrically obvious. The diagram described by Keister, Ritchie, and Washburn (Ref. 1, Appendix to Chap. 8) is convenient. Their cube has the appearance of a piece of graph paper; it is agreed that any two points which can be connected by a path made up of horizontal and vertical line segments of length 4 represent the same point on the cube. The graph paper itself has symmetries, each of which is also a symmetry of the cube. Then any two cycles which, when drawn on the graph paper cube, can be transformed into one another by a sequence of translation, reflection or rotation symmetries of the graph paper must be of the same type. Unfortunately, these symmetries account for only 128 of the 384 symmetries of the 4-cube. Cycles which are equivalent with respect to the 4-cube group 0_4 may fall into as many as three distinct types with respect to the subgroup of graph paper symmetries. This phenomenon is illustrated in Fig. 1, which shows three cycles which are of the same type in spite of the differences in their graph paper pictures.
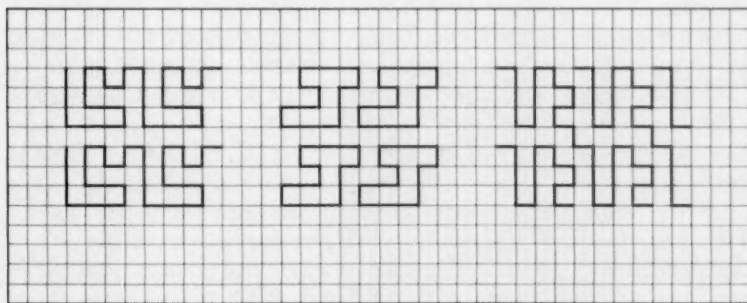

Fig. 1. — Graph paper representation of cycles on Q_4.

TABLE I — COORDINATE SEQUENCES OF CYCLES OF
DISTINCT TYPE ON Q_4

Length 2	(1 1)
Length 4	(1 2 1 2)
Length 6	(1 2 1 3 2 3)
	(1 2 3 1 2 3)
Length 8	(1 2 1 3 1 2 1 3)
	(1 2 1 3 2 4 3 4)
	(1 2 3 1 4 2 3 4)
	(1 2 3 1 4 2 4 3)
	(1 2 3 1 4 3 2 4)
	(1 2 3 4 1 2 3 4)
	(1 2 3 4 1 4 3 2)
Length 10	(1 2 1 3 4 3 1 2 1 4)
	(1 2 1 3 4 1 2 1 3 4)
	(1 2 1 3 4 1 2 1 4 3)
	(1 2 1 3 4 1 2 3 1 4)
	(1 2 1 3 4 1 2 4 1 3)
	(1 2 1 3 4 1 4 2 1 3)
	(1 2 1 3 4 1 3 1 2 4)
	(1 2 1 3 4 2 1 3 1 4)
	(1 2 1 3 4 2 1 4 1 3)
	(1 2 1 3 1 4 3 1 2 4)
Length 12	(1 2 1 3 4 3 1 2 1 3 4 3)
	(1 2 1 3 4 3 1 2 3 4 1 3)
	(1 2 1 3 4 3 1 2 3 1 4 3)
	(1 2 1 3 4 3 1 2 3 1 3 4)
	(1 2 1 3 4 3 1 3 4 2 1 3)
	(1 2 1 3 4 3 1 3 2 1 3 4)
	(1 2 1 3 4 3 1 3 2 4 1 3)
	(1 2 1 3 4 3 1 3 2 3 1 4)
	(1 2 1 3 4 1 4 2 4 1 3 4)
	(1 2 1 3 4 1 3 2 3 1 4 3)
	(1 2 1 3 4 1 4 2 4 1 4 3)
	(1 2 1 3 4 1 4 2 4 3 1 4)
	(1 2 1 3 4 2 3 1 3 4 1 3)
	(1 2 1 3 4 2 4 1 4 3 1 4)
	(1 2 1 3 4 2 1 2 3 1 2 4)
	(1 2 1 3 4 2 1 3 2 1 2 4)
	(1 2 1 3 1 2 1 4 1 3 1 4)
	(1 2 1 3 1 4 1 2 1 3 1 4)
	(1 2 3 4 1 2 4 2 3 4 2 4)
	(1 2 3 4 1 3 1 2 3 1 4 3)
	(1 2 3 4 1 3 1 2 3 4 1 3)
	(1 2 3 4 2 1 4 2 4 3 2 4)
	(1 2 3 1 4 3 1 2 3 1 4 3)
Length 14	(1 2 1 3 4 3 1 2 1 3 1 4 1 3)
	(1 2 1 3 4 3 1 2 3 4 1 4 3 4)
	(1 2 1 3 4 3 1 2 3 2 4 2 1 3)
	(1 2 1 3 4 3 1 3 4 2 4 1 3 4)
	(1 2 1 3 4 3 1 3 2 4 1 4 3 4)
	(1 2 1 3 4 1 4 2 1 4 1 3 1 4)
	(1 2 1 3 4 1 4 2 4 3 2 1 2 4)
	(1 2 1 3 4 2 1 4 2 4 3 2 1 4)
	(1 2 1 3 4 2 1 4 2 4 3 1 2 4)
	(1 2 1 3 4 2 1 3 2 3 4 2 1 3)
	(1 2 1 3 4 2 4 1 2 4 3 2 1 4)
	(1 2 1 3 4 2 4 1 2 4 3 1 2 4)
	(1 2 1 3 4 2 4 1 4 3 2 1 2 4)
	(1 2 1 3 4 2 3 2 1 2 3 4 1 3)

TABLE I — *Concluded*

	(1 2 1 3 4 2 3 1 2 3 4 2 1 3)
	(1 2 1 3 1 2 1 4 1 2 1 3 2 4)
	(1 2 1 3 1 2 1 4 1 2 3 1 2 4)
	(1 2 1 3 1 2 1 4 2 1 3 1 2 4)
	(1 2 1 3 1 2 4 1 2 1 3 1 2 4)
	(1 2 1 3 1 2 4 1 2 1 3 2 1 4)
	(1 2 1 3 2 1 2 4 1 2 1 3 1 4)
Length 16	(1 2 1 3 1 2 1 4 1 2 1 3 1 2 1 4)
	(1 2 1 3 1 2 1 4 2 1 2 3 2 1 2 4)
	(1 2 1 3 2 1 2 4 1 2 1 3 2 1 2 4)
	(1 2 1 3 2 1 2 4 2 3 2 1 3 2 3 4)
	(1 2 1 3 2 1 2 4 3 2 3 1 2 3 2 4)
	(1 2 3 2 1 2 3 4 3 2 1 2 3 2 1 4)
	(1 2 3 2 1 2 3 4 1 2 3 2 1 2 3 4)
	(1 2 3 2 1 2 3 4 1 3 1 2 1 3 1 4)
	(1 2 1 3 4 1 4 2 4 3 2 1 2 3 4 3)

A complete list of coordinate sequences for distinct types of cycles on Q_4 is given in Table I.

IV. COMPOSITION

The example (1) suggests the following result.

Theorem II. Let the L-tuple $A = (a_1, \cdots, a_L)$ and the M-tuple $B = (b_1, \cdots, b_M)$ be coordinate sequences of two paths in Q_{n-1}. Then the $(L + M + 1)$-tuple $(A, n, B) = (a_1, \cdots, a_L, n, b_1, \cdots, b_M)$ is the coordinate sequence of a path in Q_n. If each of the digits $1, \cdots, n - 1$ appears an even number of times among $a_1, \cdots, a_L, b_1, \cdots, b_M$, then (A, n, B, n) is the coordinate sequence of a cycle in Q_n.

A formal proof is easily given along the lines of the example. However, we prefer to note only that the theorem is obvious geometrically. We may envision Q_n as composed of two $(n - 1)$-cubes R and S. An extra coordinate $x_n = 0$ is added to the labels on the points of R and, similarly, $x_n = 1$ for S. Then 2^{n-1} lines are added joining corresponding points of R and S in order to construct Q_n. The path obtained by following the A path in R, then stepping over into S and following the B path is represented by (A, n, B). If all digits appear an even number of times in $a_1, \cdots, a_L, b_1, \cdots, b_M$ then the two-end points of (A, n, B) differ only in their x_n coordinate; an additional step back to R returns one to the starting point and (A, n, B, n) represents a cycle.

The n-dimensional path and cycle (if it exists) so constructed will be called the *composite path* and *composite cycle* of the paths with coordinate sequences A and B. Of the nine Hamilton lines listed, only the last is not composite.

A large family U_n of paths and cycles on the n-cube can now be constructed inductively. For $n = 2$, the family consists of all paths and cycles. When U_{n-1} has been constructed, U_n will contain all members of U_{n-1} plus all paths and cycles obtainable by making a composite from a pair of paths of U_{n-1}. It will be convenient to admit single vertices as paths (having the null coordinate sequence). Then such coordinate sequences as (n, A) with A in U_{n-1}, or even (n), represent paths in U_n. Next, we complete the construction of U_n by adding all paths and cycles equivalent to the ones just constructed. Paths and cycles belonging to U_n will be called *ultracomposite*.

The Hamilton line given as an example following Theorem I is ultracomposite, as may be seen using (1) and induction. The first five of the nine Hamilton lines in Table I are ultracomposite.

V. CHANGE NUMBERS

Let A be a coordinate sequence and let N_k, $k = 1, \cdots, n$, be the number of appearances of the digit k in A. N_k will be called the kth *change number* of A. The change numbers of any other sequence A' of the same type as A are just a rearrangment of $N_1, \cdots, N_n$. Hence a comparison of change numbers often suffices to prove two coordinate sequences to be of different type. There are, however, many examples of coordinate sequences of different type but having the same set of change numbers (the list of Hamilton lines contains two with change numbers 6, 6, 2, 2 and five with change numbers 6, 4, 4, 2).

If A and B are coordinate sequences of paths on the $(n - 1)$-cube and have change numbers $N_1, \cdots, N_{n-1}$ and $M_1, \cdots, M_{n-1}$, then the composite (A, n, B) has the change numbers

$$N_1 + M_1, \cdots, N_{n-1} + M_{n-1}, 1.$$

This observation suggests the possibility of an arithmetic test to decide whether a given set of numbers $N_1, \cdots, N_n$ are the change numbers of some ultracomposite.

In what follows we call an n-tuple $(N_1, \cdots, N_n)$ a *word* if $N_1, \cdots, N_n$ are the change numbers of an ultracomposite path. We ignore ultracomposite cycles because they may be found merely by adding single steps to those ultracomposite paths which have words in which all but one change number is even.

Following the inductive definition of U_n given above, an inductive scheme for computing all words may be given:

(i) (0,0), (0,1), (1,1) *and* (1,2) *are words.*

(ii) *If $(N_1, \cdots, N_n)$ and $(N_1', \cdots, N_n')$ are words; so are $(N_1, \cdots, N_n, 0)$ and $(N_1 + N_1', \cdots, N_n + N_n', 1)$.*

(iii) *Any permutation of the change numbers of a word produces another word.*

Thus, the only words with $n = 3$ are (000), (001), (011), (012), (111), (112), (113), (122), (123), (124), (133), and their permutations. The following theorem gives an arithmetic property of words.

Theorem III. Let the change numbers of a word be written in numerical order, $a_1 \leq a_2 \leq \cdots \leq a_n$. Then for all $k - 1, \cdots n$,

$$\sum_{i=k}^{n} a_i \leq 2^n - 2^{k-1}. \tag{2}$$

Proof: (Induction on n)

When $n = 2$, all words satisfy (2) (see (i) above). Suppose (2) is true for words of lengths $\leq n - 1$ and consider one of length n. If $a_1 = 0$, the ultracomposite path described by the given word lies entirely on a lower-dimensional face of Q_n. Then $(a_2, \cdots, a_n)$ is a word of length $n - 1$ and thus satisfies (2). If $a_1 = 1$, then the given word describes a composite path made from two paths B and C. For $i = 2, \cdots, n$, a_i is a sum $B_i + C_i$ of certain change numbers of B and C. Then for $k = 2, \cdots, n$,

$$\sum_{i=k}^{n} a_i = \sum_{i=k}^{n} B_i + \sum_{i=k}^{n} C_i \leq \sum_{i=k-1}^{n-1} b_i + \sum_{i=k-1}^{n-1} c_i, \tag{3}$$

where $b_1, \cdots, b_{n-1}$ are the numbers $B_2, \cdots, B_n$ arranged in numerical order and c_i are similarly defined. Since the b_i and c_i satisfy (2), the sums on the right of (3) total $\leq 2^n - 2^{k-1}$. Finally, when $k = 1$, (2) is necessary because there are only 2^n vertices in Q_n.

Not every n-tuple satisfying (2) is a word. An example is (0002). It is immediately recognized as not a word because every word either must contain some 1's or contain only 0's. Even ruling out such obvious examples, one still finds others, such as (114). In the case of paths of length $2^n - 1$ the following stronger result is obtained.

Theorem IV. A set of necessary and sufficient conditions that an n-tuple composed of numbers a_i, $a_1 \leq \cdots \leq a_n$ shall be the word for an ultracomposite path of length $2^n - 1$ is that (2) holds for $k = 2, \cdots, n$; that $a_1 = 1$ and, in addition, that

$$\sum_{i=1}^{n} a_i = 2^n - 1. \tag{4}$$

Proof:

Given an ultracomposite path, (2) is necessary for $k = 2, \cdots, n$ by Theorem III. Also $a_1 = 1$, since any composite has some coordinate changing only once. Equation (4) is the requirement that the length of the path be $2^n - 1$.

Conversely, suppose the given a_i satisfy the stated arithmetic conditions. If $n = 2$, the only possibility is $a_1 = 1$, $a_2 = 2$ and the coordinate sequence in question is (212). If $n \geq 3$, we are able to use a lemma of C. E. Shannon (Ref. 4, p. 84). Given a set of numbers $b_1 \leq b_2 \leq \cdots \leq b_n$ containing a pair b_r, b_s with $1 < b_r < b_s$, a *flow* operation is defined which replaces b_r by $b_r + 1$ and b_s by $b_s - 1$. Shannon shows that those sets of numbers which are obtainable by repeated flow operations starting from the initial set $1, 2, \cdots, 2^{n-1}$ are exactly those sets $a_1 \leq \cdots \leq a_n$ which satisfy our (2), (4) and $a_1 = 1$. The lemma in question states that if $1, a_2, \cdots, a_n$ is such a flow pattern then for $k = 2, \cdots, n$

$$a_k = B_k + C_k,$$

where $B_2, \cdots, B_n$ are a set of numbers (not necessarily in increasing order) obtainable from $1, 2, \cdots, 2^{n-2}$ by means of flows, and similarly for $C_2, \cdots, C_n$. By induction, $(B_2, \cdots, B_n)$ and $(C_2, \cdots, C_n)$ are themselves words for some paths B and C. Then $(B, 1, C)$ is a path having the word $(a_1, \cdots, a_n)$ and the theorem is proved.

The decomposition of a_k into a sum $B_k + C_k$ can be done either by inspection or using the procedure by which Shannon's lemma is proved. Thus the problem of finding a path for a given word reduces to two such problems in lower-dimensional cubes. Continuing, one finally requires only 2^{n-2} paths on certain square faces of Q_n. Since the decomposition of the a_k can generally be done in several different ways, there may be many types of paths which can be found for a given word.

VI. HIGHER DIMENSIONS

For the two applications mentioned earlier, one might suppose that a complete list of all types of paths, or only of Hamilton lines, would be useful. Such tables must lengthen rapidly as n increases. For example, the number H_n of types of ultracomposite Hamilton lines is at least as large as the number of words, satisfying Theorem IV, for which $a_2, \cdots, a_n$ are all even. An enumeration of such words shows $H_5 \geq 19$. A rapidly increasing lower bound on H_n is now given.

Theorem V. Let $s_0, s_1, \cdots$ be defined by the recurrence

$$s_{n+1} = (n + 1)s_n - \binom{n}{2} s_{n-2} \tag{5}$$

and the initial values $s_0 = 1$, $s_1 = 1$, $s_2 = 2$. *Then* $H_{n+2} \geqq s_n$. *For large n, an asymptotic formula is*

$$s_n \approx n!(n\pi)^{-(1/2)}e^{3/4}. \tag{6}$$

Proof:

By Theorem IV the $(n + 1)$-tuple $W = (2^1, 2^2, \cdots, 2^n, 1)$ is a word for some coordinate sequence T. Let a permutation P, say $k \to p(k)$, be applied just to the first n coordinate. Then T changes to PT with the word $(2^{p(1)}, \cdots, 2^{p(n)}, 1)$. The coordinate sequence $S(P) = (T, n + 2, PT, n + 2)$ describes an ultracomposite Hamilton line in which the kth coordinate change number is

$$N_k(P) = \begin{cases} 2^k + 2^{p(k)} & \text{if } k = 1, \cdots, n \\ 2 & \text{if } k = n + 1 \text{ or } n + 2. \end{cases}$$

Among the $n!$ Hamilton lines $S(P)$ there are at least as many distinct types as there are distinct sets of coordinate change numbers. It follows from the uniqueness of binary notation that $2^a + 2^b = 2^c + 2^d$ is equivalent to the statement that the (unordered) pairs (a, b) and (c, d) are the same. If one of the coordinate change numbers $2^K + 2^{p(K)}$ of $S(P)$ equals a coordinate change number of some other line $S(P')$ then P' [say $i \to p'(i)$] satisfies either $p'(K) = p(K)$ or $p'[p(K)] = K$. Then, in order for P' to have the same set of coordinate change numbers as P, every cycle of P' must either be a cycle of P or the inverse of a cycle of P. To get a lower bound on H_{n+2}, we may count the number s_n of equivalence classes of permutations in which P and P' are considered equivalent when every cycle of P' is either a cycle of P or the inverse of a cycle of P. Precisely this enumeration is also required for Cayley's problem of counting the number of terms in the expansion of a symmetric determinant. In this connection, (5) and (6) were given by I. Schur. Derivations may be found in Pólya and Szégö's book (Ref. 3, Vol. 2, Ch. 7, probs. 45 and 46).

In higher-dimensional cubes, it seems likely that the majority of the types of Hamilton lines will not be ultracomposite or even composite. To support this guess, we now construct a large class of non-composite lines. The last of the nine types listed for $n = 4$ was one such type. Using its coordinate sequence, $(1213 \cdots 3)$, we construct sequences of the form

$$A = (1, A_1, 2, A_2, 1, A_3, \cdots, 3, A_{16}),$$

where each of $A_1, \cdots, A_{16}$ represents a path of length $2^{n-4} - 1$ and contains only the digits $5, \cdots, n$. We also require that $5, \cdots, n$ each

appear an even number of times. Each such A is the coordinate sequence of a non-composite Hamilton line.

REFERENCES

1. Keister, W., Ritchie, A. E. and Washburn, S. H., *The Design of Switching Circuits*, D. Van Nostrand and Co., N. Y., 1951.
2. König, D., *Theorie der Endlichen und Unendlichen Graphen*, Akad. Verlag., 1936, reprinted Chelsea Publ., N. Y., 1950.
3. Pólya, G. and Szégö, *Aufgaben und Lehrsatze aus der Analysis*, Dover Publ. N. Y., 1945.
4. Shannon, C. E., The Synthesis of Two-Terminal Switching Circuits, B.S.T.J., **28,** January, 1949, pp. 59-98.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

AMBROSE, J. F., see McCALL, D. W.

ANDERSON, P. W.

The Reaction Field and Its Use in Some Solid-State Amplifiers, Monograph 2910.

AUGUSTYNIAK, W. M., see BEMSKI, G.

BATTERMAN, B. W.

Hillocks, Pits, and Etch Rate in Germanium Crystals, Monograph 2911.

BELLOWS, B. C. and GRAHAM, R. S.

Experimental Transversal Equalizer for TD-2 Radio Relay System, Monograph 2903.

BEMSKI, G. and AUGUSTYNIAK, W. M.

Annealing of Electron Bombardment Damage in Silicon Crystals, Monograph 2912.

BOBECK, A. H.

New Storage Element for Large-Sized Memory Arrays — the Twistor, Monograph 2904.

BRIDGERS, H. E., see THEUERER, H. C.

BUEHLER, E., see THEUERER, H. C.

CHYNOWETH, A. G. and McKAY, K. G.

Threshold Energy for Electron-Hole Pair-Production by Electrons in Silicon, Monograph 2913.

CLOGSTON, A. M., see GESCHWIND, S.

COOK, J. S., KOMPFNER, R. and YOCUM, W. H.
Slalom Focusing, Monograph 2914.

DESOER, C. A.
Network with Periodically Operated Switch Solved by Successive
Approximations, Monograph 2905.

DUNN, H. K., see YOUNG, R. W.

FOSTER, F. G., see WILLIAMS, H. J.

GESCHWIND, S. and CLOGSTON, A. M.
Narrowing Effect of Dipole Forces on Inhomogeneously Broadened
Lines, Monograph 2915.

GRAHAM, R. S., see BELLOWS, B. C.

GYORGY, E. M.
Rotational Model of Flux Reversal in Square Loop Ferrites, Mono-
graph 2901.

HAMMING, R. W.
Controlling the Digital Computer, Monograph 2916.

HARKER, K. J.
Use of Scanning Slits for Current Distribution in Electron Beams,
Monograph 2917.

HUTSON, A. R.
Hall Effect Studies of Doped Zinc Oxide Single Crystals, Monograph
2918.

KELLY, E. M., see WILLIAMS, H. J.

KOMPFNER, R., see COOK, J. S.

KRONACHER, G.
Design, Performance and Application of the Vernier Resolver,
Monograph 2906.

LANZA, V. L., see McCALL, D. W.

LAWSON, C. C.

Buried Distribution for Telephone Circuits, Monograph 2919.

MacCALL, L. A.

Theory of the Relativistic Oscillator, Monograph 2920.

MASON, W. P. and THURSTON, R. N.

Piezoresistive Materials in Measuring Displacement, Force and Torque, Monograph 2924.

McCALL, D. W., AMBROSE, J. F. and LANZA, V. L.

Diffusion in Ethylene Polymers, Monograph 2921.

McCALL, D. W. and SLICHTER, W. P.

Molecular Motion in Polyethylene, Monograph 2922.

McKAY, K. G., see CHYNOWETH, A. G.

McSKIMIN, H. J.

Ultrasonic Propagation in Liquids, Monograph 2923.

MERTZ, P. and MITCHELL, D.

Data Transmission Service Using Private Line Voice Telephone Channels, Monograph 2907.

MITCHELL, D., see MERTZ, P.

OLSEN, K. M.

Superpurity Nickel Melted Under Controlled Atmospheres, Monograph 2925.

PRIM, R. C.

Shortest Connection Networks and Some Generalizations, Monograph 2908.

SHERWOOD, R. C., see WILLIAMS, H. J.

SLICHTER, W. P., see McCALL, D. W.

SUHL, H.

Theory of the Ferromagnetic Microwave Amplifier, Monograph 2926.

THEUERER, H. C., WHELAN, J. M., BRIDGERS, H. E. and BUEHLER, E.

Heat Treatment of Silicon Using Zone-Heating Techniques, Monograph 2927.

THURSTON, R. N., see MASON, W. P.

ULRICH, W.

Non-binary Error Correction Codes, Monograph 2909.

WAHL, A. J.

Three-Dimensional Analytic Solution for Alpha of Alloy Junction Transistors, Monograph 2928.

WHELAN, J. M., see THEUERER, H. C.

WILLIAMS, H. J., SHERWOOD, R. C., FOSTER, F. G. and KELLY, E. M.

Magnetic Writing on Thin Films of MnBi, Monograph 2929.

WILLIAMS, R. C.

Radiation Due to Adhesive Breaks Between High Polymers and Glass, Monograph 2902.

YOCUM, W. H., see COOK, J. S.

YOUNG, R. W. and DUNN, H. K.

On the Interpretation of Certain Sound Spectra of Musical Instruments, Monograph 2930.

Contributors to This Issue

GERHARD BACKENSTOSS, Dipl. Phys., 1949, Dr. rer. nat., 1953 University of Freiburg i.Br. (Germany); research associate in the Institute of Physics, University of Freiburg i.Br. (Germany), 1953–55; Bell Telephone Laboratories, 1955–57. At Bell Laboratories Mr. Backenstoss was associated with studies of silicon transistors. He was especially concerned with investigation of the diffusion of impurities into silicon; measurement of mobility of holes and electrons in heavily doped silicon, and with studies of changes of minority carrier lifetime by electron bombardment. Member German Physical Society.

CURTIS L. BEATTIE, M.A., 1933, University of Georgia; Graduate studies, Johns Hopkins University, 1946, and Columbia University, 1947; acting head of mathematics department, Mercer University, 1941–42; U. S. Army Signal Corps Engineering Laboratory, 1942–45; head of mathematics department, Judson College, 1946–47; assistant professor of mathematics, University of Bridgeport, 1947–53; Bell Telephone Laboratories, 1954–. Mrs. Beattie has worked on analysis and programming of mathematical problems relating to radio and waveguide research. Member American Mathematical Society.

HOWARD BOYET, B.S., 1954, College of the City of New York; Ph.D., 1953, New York University; National Advisory Committee for Aeronautics, 1944–47; New York University, faculty, 1948–53 and research assistant, Institute for Mathematical Sciences, 1952–53; Bell Telephone Laboratories, 1953–57. With Bell Laboratories, Mr. Boyet was engaged in solid-state device development in the field of microwave ferrite devices, including ferrite scanning antennas, isolators and circulators. Member American Physical Society, American Institute of Physics.

THOMAS F. EGAN, Bell Telephone Laboratories, 1930–. Mr. Egan has specialized in analytical chemistry, including the application of microchemical analysis to engineering problems. From 1938 to 1944 he engaged in chemical diagnoses of central office problems relating to noise and contact resistance. Since 1945 he has concentrated on problems of dust in central office apparatus, especially in relation to open-contact failures

831

and dust exclusion in relay cabinet design. Member American Chemical Society.

E. N. GILBERT, B.S., 1943, Queens College; Ph.D., 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–46; Bell Telephone Laboratories, 1948–. Mr. Gilbert has been engaged in studies of information theory and switching theory. He is a member of the Communication Fundamentals Group. Recipient M.I.T. Applied Mathematics Fellowship, 1946–48. Member American Mathematical Society.

ROBERT H. GUMLEY, B.E.E., 1939, New York University; Bell Telephone Laboratories, 1930–. Mr. Gumley was first engaged in apparatus wiring and adjusting including work on the No. 1 crossbar system. He later worked on relay requirements and testing and during World War II on design and testing of operation flight trainers. Since 1945 he has been engaged in relay performance studies and relay engineering. He is now head of a group studying relay contact problems. Member Eta Kappa Nu.

H. W. HERMANCE, Crucible Steel Co., 1917–19; Procter and Gamble 1921-25; Western Electric Co., 1925–27; Bell Telephone Laboratories 1927–. After working on analytical control of materials with Western Electric, Mr. Hermance joined the Laboratories, where he has specialized in microanalytical methods. He has had a prominent part in developing analytical techniques and microchemical laboratory facilities. Since 1945 he has been in charge of a group specializing in the diagnosis of chemical and related problems as they affect switching apparatus in the field. Member American Association for the Advancement of Science, American Chemical Society, Metropolitan Microchemical Society.

H. J. KEEFER, B.S. in E.E., 1937, Georgia School of Technology; Bell Telephone Laboratories, 1928–. As a member of technical staff Mr. Keefer was engaged in laboratory testing of step-by-step system PBXs and long-line pulsing circuits until World War II, when he served as an Artillery and Signal Officer. Since his return to the Laboratories, Mr. Keefer has concentrated on relay contact studies.

ALLEN H. MEITZLER, B.S., 1951, Muhlenberg College; M.S., 1953, and Ph.D., 1955, Lehigh University; Bell Telephone Laboratories, 1955–. As a member of the Military Apparatus Development Department, Mr.

Meitzler has been concerned with the development of ultrasonic solid delay lines. Recipient of Hood Fellowship, Lehigh University, 1955. Member I.R.E., American Physical Society, Acoustical Society of America, Sigma Xi.

STEPHEN O. RICE, B.S., 1929, Oregon State College; graduate studies California Institute of Technology, 1929–30 and 1934–35; Bell Telephone Laboratories, 1930–. In his first years at the Laboratories, Mr. Rice was concerned with nonlinear circuit theory, especially with methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigation of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He is a Gordon McKay Visiting Lecturer in applied physics at Harvard University for the Spring 1958 term. Fellow I.R.E.

C. F. P. ROSE, Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Mr. Rose has been concerned with design, development, installation and testing of short-wave transoceanic transmitters in the United States and South America. His recent work with the Radio Research Department has been studies of millimeter waves and development of techniques for constructing helix waveguides for microwave transmission. Senior member I.R.E.

HAROLD SEIDEL, B.E.E., 1943, College of the City of New York; M.E.E., 1947, and D.E.E., 1954, Polytechnic Institute of Brooklyn; Microwave Research Institute of Polytechnic Institute of Brooklyn, 1947; Arma Corp., 1947–48; Federal Telecommunications Labs, 1948–53; Bell Telephone Laboratories, 1953–. He has been concerned with general electromagnetic problems, especially regarding waveguide applications, and with analysis of microwave ferrite devices. Member I.R.E., Sigma Xi.

FRIEDOLF M. SMITS, Dipl. Phys., Dr. rer. nat., 1950, University of Freiburg i.Br. (Germany); research associate, Physikalisches Institut, University of Freiburg i.Br. (Germany), 1950–54; Bell Telephone Laboratories, 1954–. Mr. Smits was first concerned with studies of impurity diffusion in germanium and silicon for semiconductor device applications. He is now engaged in exploratory studies of semiconductor devices. Member American Physical Society, German Physical Society.

H. L. Stadler, A.B., 1948, Harvard College; S.M., 1951, and Ph.D. 1954, University of Chicago; Bell Telephone Laboratories, 1954–. Mr. Stadler has been engaged in development of barium titanate ferroelectric devices. Member American Physical Society, American Institute of Physics, Sigma Xi.

William D. Warters, A.B., 1949, Harvard College; M.S., 1950, and Ph.D., 1953, California Institute of Technology; Bell Telephone Laboratories, 1953–. Mr. Warters has concentrated on studies of circular waveguides for long-distance, broadband transmission. He was the recipient of a National Science Foundation Predoctoral Fellowship in 1952. Member American Physical Society, Sigma Xi, Phi Beta Kappa.